REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
28-08-2012	Final Report	19-Jan-2011 - 18-Jul-2011

4. TITLE AND SUBTITLE
Robust and Efficient Anti-Phishing Techniques

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W911NF-11-C-0046

5c. PROGRAM ELEMENT NUMBER
606055

6. AUTHORS
Nina Kohli-Laven, John F. Buford

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Altusys Corporation at Princeton
P O Box 1274

Princeton, NJ 08542 -1274

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58869-CS-SB1.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
The OSD needs high-accuracy and low-latency automatic identification and mitigation techniques to detect and stop phishing attacks. This project researched and developed socio-linguistic indicators that can be used to support more fine-grained, accurate detection of phishing emails. The indicators address several different types of phishing emails, including social malware emails, and demonstrate the feasibility and desirability of adding socio-linguistic attributes to phishing detection signatures. The project focused on : feasibility study of socio-linguistic attributes

15. SUBJECT TERMS
phishing, anti-phishing, socio-linguistic analysis, heuristic classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Khushboo Shah
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 609-651-4500

Report Title

Robust and Efficient Anti-Phishing Techniques

ABSTRACT

The OSD needs high-accuracy and low-latency automatic identification and mitigation techniques to detect and stop phishing attacks. This project researched and developed socio-linguistic indicators that can be used to support more fine-grained, accurate detection of phishing emails. The indicators address several different types of phishing emails, including social malware emails, and demonstrate the feasibility and desirability of adding socio-linguistic attributes to phishing detection signatures. The project focused on : feasibility study of socio-linguistic attributes as indicators of phishing; development of socio-linguistic features for detection of phishing emails; development of a prototype for collecting socio-linguistic features from phishing emails.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Paper

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Post Doctorates

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Faculty Supported

NAME	PERCENT_SUPPORTED

FTE Equivalent:
Total Number:

Names of Under Graduate students supported

NAME	PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

The final report is uploaded as an attachment.

Technology Transfer

Final Report

For Contract: W911NF-11-C-0046

COR: Dr. Cliff Wang

Robust and Efficient Anti-Phishing Techniques

August 27, 2012

Prepared by:

Altusys Corp.

P. O. Box 1274, Princeton, NJ 084542

609-651-4500

www.altusystems.com

Table of Contents

Table of Figures

1. Summary

1.1 Objective

Research and develop socio-linguistic indicators that can be used to support more fine-grained, accurate detection of phishing emails. The indicators should address several different types of phishing emails, including social malware emails, and should

demonstrate the feasibility and desirability of adding socio-linguistic attributes to phishing detection signatures.

1.2 Description

The OSD needs high-accuracy and low-latency automatic identification and mitigation techniques to detect and stop phishing attacks. Phishing has evolved from a nuisance into a top security concern. As the number, cost, and complexity of phishing attacks continue to increase, robust and effective techniques are critically needed to counter the new threats. Existing solutions such as spam filters rely heavily on manually maintained blacklists of phishing websites, and are not robust at catching phishing emails, especially spear-phishing attacks, since these attacks look just like legitimate emails.

Altusys has investigated how to apply socio-behavioral analysis, specifically analysis of the linguistic patterns in phishing and legitimate emails, to the detection of phishing emails.

The first goal of this study is to investigate the feasibility of applying socio-linguistic analysis to phishing email detection. A major focus is on establishing whether the socio-linguistic characteristics of emails are different between phishing and legitimate emails. Another focus is on established whether emails from the same author have a consistent socio-linguistic "signature" that can be used to detect inauthentic emails (e.g. social malware) sent from that author's email account.

The second goal is to develop and describe socio-linguistic features of phishing emails such that they can be incorporated into detection algorithms for a phishing detection system.

The techniques developed under this research will result in enhanced phishing email detection. The project leverages the analytical power of socio-linguistics in order to enhance automated analysis of email. It provides novel instruments for automating this analysis: socio-linguistic features can be quantified and incorporated into phishing detection systems.

2. Project Milestones and Deliverables

Per agreement with the COR at the kickoff meeting, **milestones were modified to reflect COR focus on sociolinguistic anti-phishing research.** The following sub-sections summarize each milestone and deliverable.

The Phase 1 project focused on:

1. Feasibility study of socio-linguistic attributes as indicators of phishing
2. Development of socio-linguistic features for detection of phishing emails
3. Development of a prototype for collecting socio-linguistic features from phishing emails

2.1 Establish Feasibility of using Socio-Linguistic Attributes to Detect Phishing

Milestone: By month 3

Status: Completed.

2.2 Develop and Refine Socio-Linguistic Phishing Detection Features

Milestone: By month 5

Status: Completed

2.3 HPPS Prototype

Milestone: Implement HPPS email analyzer based on the socio-linguistic models developed earlier

Status: Completed

2.4 Summary of Deliverables by Milestone

The following table lists each deliverable by milestone. Each deliverable is attached to this report in the indicated appendix.

Table 1 Deliverables by Milestone

Milestone	Delivered	Deliverable	Appendix
Establish feasibility of using socio-linguistic attributes to detect Phishing	2/19/11 3/19/11	Socio-Linguistic Anti-Phishing Feasibility Study Draft 1 Final Report	10.1
Develop and Refine Socio-Linguistic Phishing Detection Features	4/19/11 5/19/11 6/19/11 7/19/11 7/19/2011 5/19/11 7/19/11	Socio-Linguistic Detection Features of Phishing Emails Documentation: Draft 1 Draft 2 Draft 3 Final Report Grading the 8 Socio-Linguistic Features of Phishing Emails Attack Model: Social Malware Report Draft 1 Final Report	10.2 10.3 10.4 10.5
HPPS Prototype	3/19/2011 3/19/2011 8/27/2012 8/27/2012	Requirements Architecture Prototype Testing	3.7, 10.6 3.8 3.9 3.10

3. Summary of Technical Activity

3.1 Overview

The problem addressed in Phase I is to develop socio-linguistic features that can support low latency, low false positive, low false negative phishing email detection in a heuristics-based protection system.

3.2 Characterize Phishing Email Problem

Altusys investigated variation and developed a typology of phishing emails from a socio-linguistic perspective.

There are two principal types of phishing email, each requiring a different classification strategy, for the purposes of socio-linguistic analysis.

Type 1 are **emails pretending to be official communication from trusted institutions or businesses.** The emails usually pose as official communication from a financial or online payments or retail company or institution and dupe recipients into visiting and providing personal financial information.

Type 2 are **emails pretending to be communication from contacts whom the email recipient knows** as part of his or her social and/or professional world. For example, emails posing to be authored by co-workers, family members, or contacts from alumni associations, neighborhood associations, or social networks in online communities, that are familiar to the recipient. The emails use social engineering to craft a message that the recipient is more likely to trust because it has the appearance of communication from a friend, colleague, family member or associate. The emails can involve a range of different content types and topics.

3.3 Establish Feasibility of Using Socio-Linguistic Attributes to Detect Phishing

Altusys formulated the following hypotheses in order to initiate the research and development of socio-linguistic indicators:

Hypothesis # 1: Type 1 Phishing emails possess certain unique socio-linguistic features. Type 1 Phishing emails can be identified as Phishing when analyzed for the presence or co-presence of these socio-linguistic features.

Hypothesis # 2: Email authors have distinct styles ("idiolects") that can be identified by analyzing their socio-linguistic features. Type 2 emails can be identified as Phishing when analyzed for the presence or lack of presence of socio-linguistic features that have been identified as the signature style, idiolect, of a particular author.

Altusys confirmed the hypotheses by conducting a preliminary analysis of its dataset of phishing emails. A summary of the report on this research is in Appendix 10.1.

The Type 1/Hypothesis 1 dataset contains 196 Phishing emails from 25+ institutions (including banks, online payments, online retailers, and social networking services such as Fidelity, Twitter, and iTunes, Amazon, Paypal, Ebay, Visa, Facebook, Bank of America, Chase, and Citibank.) For a description of the dataset, including the legitimate email dataset used for comparison, by vendor-type, see Table 2. We eliminated duplicate or highly similar phishing emails from the same vendor from the dataset in order to cover the broadest possible range of phishing email types and focused on phishing emails sent 2008 and after.

Table 2 Type 1 Phishing Email Dataset

	Vendor	Type	No. of Emails
1	Chase	Bank	56
2	Royal Bank of Canada	Bank	48
3	Bank of America	Bank	2
4	Other Assorted Banks	Bank	8
5	Amazon	Online Retailer	1
6	AOL	Online Service	5
7	UPS	Online Service	1
8	eBay	Online Service	19
9	Facebook	Social Networking Service	2
10	PayPal	Online Service	27
11	Visa	Online Service	12
12	Fidelity	Financial Services	1
13	iTunes	Online Retailer	1
14	Other	Online Services/Retailer	13

Table 3 Type 1 Legitimate Email Dataset

	Vendor	Type	No. of Emails
1	Citibank	Bank	7
2	Chase	Bank	2
3	Royal Bank of Canada	Bank	1
4	Mint	Online Service	1
5	Yahoo	Online Service	1
6	Skype	Online Service	1
7	Amazon	Online Retailer	1

The Type 2/Hypothesis 2 dataset contains 10 social malware Phishing emails (e-card, social networking invitations, job offer, distress emails) and, in order to profile legitimate emails, we selected emails from 6 authors in the Enron email corpus. Our Enron test set contained about 600 emails written in 2001.[1]

Key finding # 1: For Type 1 emails: Semantic and pragmatic linguistic features of phishing emails identified in the proposal and kick-off are valid differentiators of phishing and legitimate emails. Differences in basic structural and stylistic patterns, such as spelling, word counts, punctuation use, and spacing, are also present, though may not be differentiators when considered in isolation.

Altusys Recommendation: In order to capture the broadest range of socio-linguistic indicators of Phishing: **focus on developing an integrated set of socio-linguistic indicators of phishing that reflect both semantic/pragmatic dimensions of phishing text, and simpler linguistic differences such as spelling, word counts, punctuation, etc.**

Key finding # 2: For Type 2 emails: Although semantic and pragmatic differences in the meaning of language used in emails provided cues to author identity, simple linguistic patterns would suffice as a differentiator of authors. (Examples of simpler patterns are lexical choices, as well as format, length, signature, punctuation, and basic syntactical choices e.g. full sentences with pronouns, prepositions, verbs and nouns vs. sentences that cut-off personal pronouns etc.).

Altusys Recommendation: Focus on simpler patterns rather than complex semantic and pragmatic analysis when profiling author email signatures. Focusing on these simpler linguistic features instead of more complex semantic or pragmatic patterns is also attractive because even short emails (1-line or 2-line) can be rich with data about these simpler patterns.

3.4 Develop Socio-Linguistic Indicators of Phishing Emails

For Type 1 Phishing emails, Altusys compared Type 1 Phishing emails to legitimate institutional email communication to generate and validate the list of features. The features identified, described, and refined are listed in Table 4. A more detailed report on the features, including the refinements described in Section 3.5, is in Appendix 11.2. Examples of Phishing emails diagrammed using the features (Fig. 1) is in Appendix 11.3.

Figure 1 shows a sample of diagrammed email from the test set using the features identified and refined in Months 1-5. Left: Phishing email posing as official communication from Amazon (2010). Right: Legitimate email from Citibank (2010). See Appendix 11.3.

[1] Public access to the Enron email corpus provided by ZL Technologies at: http://edrm.net/resources/data-sets/edrm-enron-email-data-set-v2 .

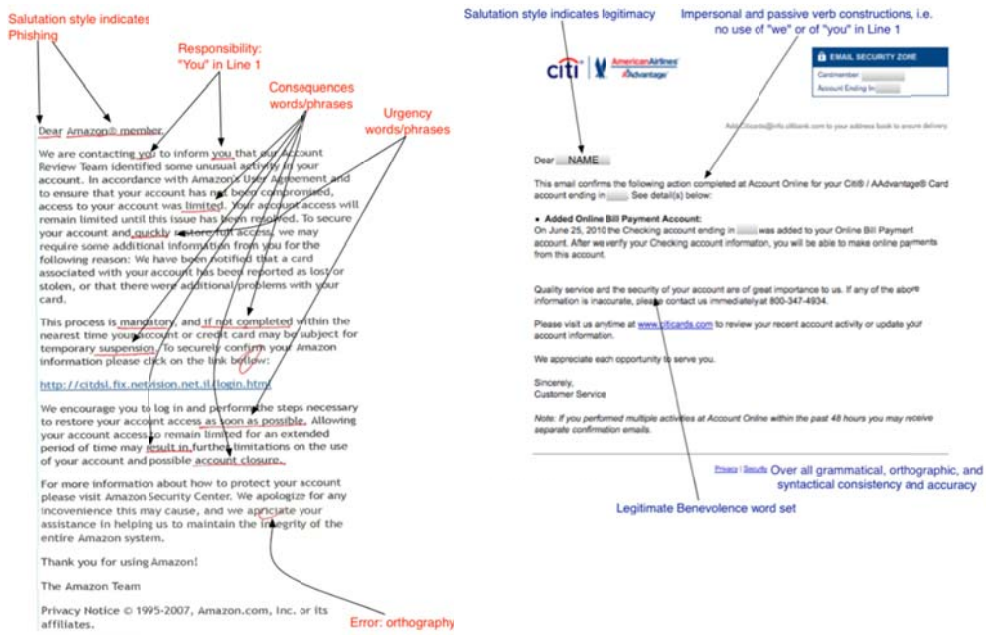


Figure 1 Sample of diagrammed emails

Table 4 Overview of Identified Socio-Linguistic Features of Phishing Emails

	Feature	Description	Characteristics
1	**Consequences**	How the Phisher makes the message seem consequential (pragmatic type)	1. Use of subjunctive constructions 2. Listing of consequences of not acting, e.g. "if you do not respond…" "account closure"
2	**Urgency**	How the Phisher makes the message and any response to it urgent and imperative (pragmatic type)	1. Qualifying time with words such as "now" and "immediately" 2. Using constructions that imply time sensitivity such as "Alert" 3. Using imperative verb constructions (e.g. Stop, go, login, click), the deontic modal construction (e.g. "you must"), and other regular verbs expressing necessity (e.g. "need")
3	**Errors**	Phishing emails have a higher rate of misspellings (orthographic errors) and grammatical error (simple type)	Most common observed errors: 1. Non-agreement of subject/verb 2. Misplaced infinitives 3. Erroneous pluralization 4. Orthographic error (spelling, punctuation)
4	**Benevolence**	How the Phisher makes the message seem in the	Phrases implying the virtuosity, concern, or diligence of the sender:

		reader's best interests	1. "for your protection" 2. "bring to your attention" 3. "Valued customer"
5	**Authority**	How the Phisher makes the message seem authoritative	1. Strategic use of the 'pointing; pronouns "we," "us," and "our" 2. Allusion to the official nature of the communication
6	**Responsibility**	How the Phisher makes it seem it is incumbent on the recipient to personally take action	Strategic and frequent use of the 'pointing; pronouns "you" and "your."
7	**Salutation**	Phishing emails initiate the communication differently	More frequent use of "Dear" to address the recipient
8	**Tense**	Tense choice is different in Phishing emails	More frequent use gerunds and relative tenses

For Type 2 Phishing emails, Altusys examined social malware emails and compared them to legitimate social emails in the Enron dataset to generate a list of socio-linguistic features. In the case of generic social malware spam emailed en masse and promising links to photos, news, or profiles of friends, some of the same principles used to detect Type 1 Phishing emails (described in Table 4) apply. Responsibility and Authority are common socio-linguistic features of these emails. These emails tend to have distinct patterns of Salutation. They also commonly use specific types of Urgency language – specifying time frames in which the recipient needs to act on the information in the email. Figure 2 shows an E-card Phishing email diagrammed using same or similar socio-linguistic diagnostic categories as in the case of institutional Phishing emails

In the case of emails targeted to prominent public or business figures that were crafted based on the particular interests and networks of those individuals, Altusys chose to model attack methods in order to identify the socio-linguistic mechanisms used to produce these emails (See Section 3.6).

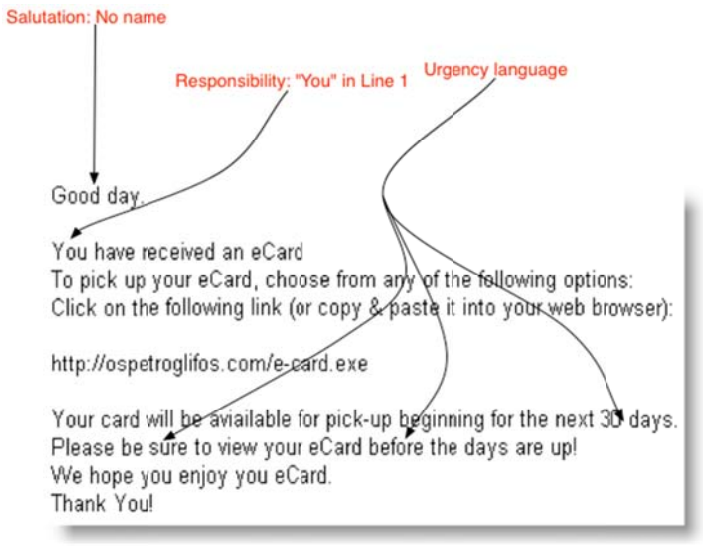


Figure 2 Diagrammed E-card Phishing email

3.5 Refine Socio-Linguistic Indicators of Phishing Emails

Altusys added the following refinements the socio-linguistic features of Phishing emails:

Grading of Features: Altusys split the recommended rules for determining whether a feature is present into grades. Grade 1 are rules that strongly match the templates used in Phishing emails. Grade 2 are rules that match the language templates used in Phishing emails but that also match some language templates used in Legitimate emails (i.e. weaker diagnostic of Phishing). For a complete list of Grades for each of the 8 features, see Appendix 11.4.

Scoring: The probability that email is Phishing when it contains the features is described for each feature and for combinations of features in Table 5.

Table 5 Phishing Probability for Co-Presence of Various Themes

Feature	Consequences	Urgency	Errors	Benevolence	Authority	Responsibility	Salutation	Tense
Consequences	Med/High	Med/High	High	Med/High	Med/High	Med/High	High	High
Urgency	Med/High	Med/High	High	Med/High	Med/High	Med/High	High	High
Errors	High	High	High	High	High	High	High	High
Benevolence	Med/High	Med/High	High	**Low**	Medium	Medium	High	High
Authority	Med/High	Med/High	High	Medium	Medium	Medium	High	High
Responsibility	Med/High	Med/High	High	Medium	Medium	Medium	High	High
Salutation	High	High	High	High	High	High	High	High
Tense	High	High	High	High	High	High	High	High

3.6 Elaborate Attack Models for Social Malware Phishing

Unlike Type 1 emails, each socially engineered phishing email may be unique, especially in the case of emails targeted to prominent public or business figures that were crafted based on the particular interests and networks of those individuals. There is a range in sophistication of how socially engineered emails are crafted for their targeted audience,

with some Phishers using generic templates to target members of social networks (e.g. standard e-card emails sent out on major public holidays) while others craft specific emails to match their victim's interests (e.g. soccer news emails for web users who visit soccer sites).

3.7 HPPS Requirements

Altusys defined 90+ requirements for the HPPS (Hybrid Phishing Protection System). The requirements document is in Appendix 11.6 of this document.

3.8 HPPS Architecture

Altusys defined an architecture document for HPPS.

3.9 HPPS Prototype

Altusys developed a prototype HPPS which consists of:

- An add-in for Microsoft Outlook 2010 which allows the user to select an email for processing
- A set of HPPS text processing services running as extensions to a semantic wiki [9], specifically using SMW+ [10]
- An integration with the Proxem Antelope (Advanced Natural Language Object-oriented Processing Environment) [5]

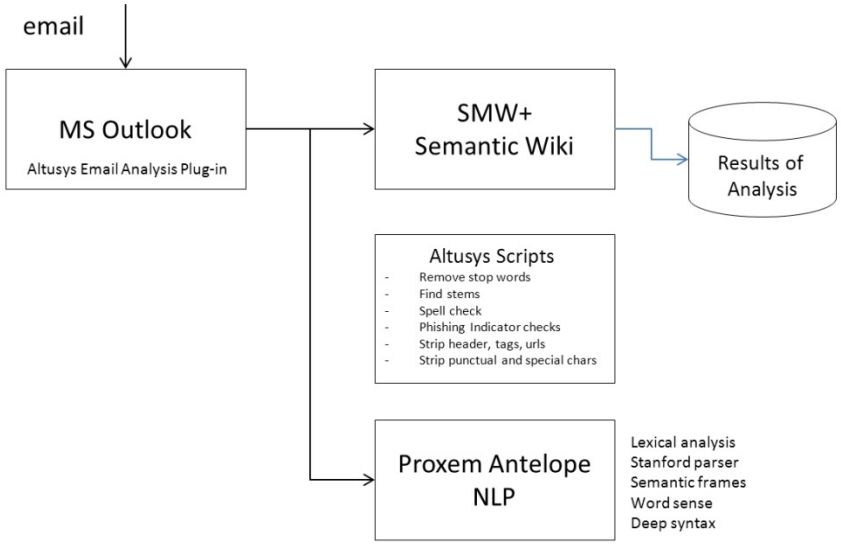


Figure 3 Architecture of prototype system

A screen shot of the HPPS add-in to Microsoft Outlook 2010 is shown in Figure 4. The user first selects a specific email and then selects the "Import and Analyze Email" menu option. The email will be processed and the resulting analysis is automatically inserted into the semantic wiki. The processing steps include:

- Remove the email header
- Remove HTML markup

- Remove special characters
- Remove stop words
- Count word frequency
- Count word stem frequency
- Count misspellings
- Count use of **authority** terms from the following list: federal reserve, federal government, federal, irs, internal revenue service, treasury department, fbi, federal bureau of investigation, dhs, homeland security, senate, army, navy, air force, mint, fort knox, bill, invoice, password, account, certify, certificate, audit, tax, president, minister, prime minister, senator, governor, general, colonel, sgt, seargant, lieutenant, chief of staff, king, advisor, ceo, coo, attorney, barrister, doctor, executive, officer, official, officials, manager, administrator, admin, system administrator, officers, director, board member, secretary, professor, prof, auditor, vp, vice president, world bank, international monetary fund, imf, central bank, united nations, america, usa, chase, citibank, jpmorgan, bank of america, bofa, hsbc, ubs, trust, probate, court, judge, his honor, honorable, justice, lloyds, first direct, bank of england, ups, usps, paypal, visa, mastercard, amex, american express, fidelity, bank, ebay, ibm, amazon, google, yahoo, microsoft, royal bank, swiss bank, numbered account, fedex, dhl, ups, wall st, wall street, fortune 100, fortune 500, olympic, olympics, nyse, gold, silver, platinum, stock exchange, rolex, tiffany, represent, representative, aol, past due, facebook, western union, guarantee, government, division, senior, overseas, highly placed, security companies, security company, capital, top secret, clearance, partner, transaction, contract, legal, law, god bless, business proposal, metlife, insurance, wire transfer, money gram, moneygram, lottery, confidential, consulate, ambassador, diplomat, diplomatic, embassy, consul, adobe, linkedin, twitter, skype, intuit, award, reward, records
- Count use of **urgency** terms from the following list: immediate, immediately, hurry, soon, delay, now, certain, sure, late, need, alert , attention, expire, expires, expired, expiration, quickly, quick, fast, must, chance, opportunity, act, today, asap, final, notice, strict, strictly, death, illness, urgent, urgency, running, important, dead, obligatory, require, requires, requirement, required
- Count use of **consequences** terms/phrases from the following list: if you do not respond, failure to respond, result in, lead to, account closure, penalty, rejected, last chance, termination, suspend, failure, deactivate, deactivated, block, blocked, closed, close, deleted, lock, locked, expired
- Count use of **benevolence** terms/phrases from the following list: for your protection, for your benefit, bring to your attention, valued customer, friend, friendship, personal, mutual benefit, best interest, worry about, winner, apologize
- Determine the **salutation** used
- Determine if **subjunctive mood** was used in any sentences

- Invoke the Proxem antelope service to obtain, for each sentence, parse tree, word usage, and semantic frames

All results are placed in the semantic wiki and the summary table (Figure 5) is automatically updated.

Figure 5 shows the HPPS Analysis and Data Collection Wiki. The main email analysis page shows the summary of each email that is processed by the prototype. More details for each email can be found on the specific page for that email, by following the associated link.

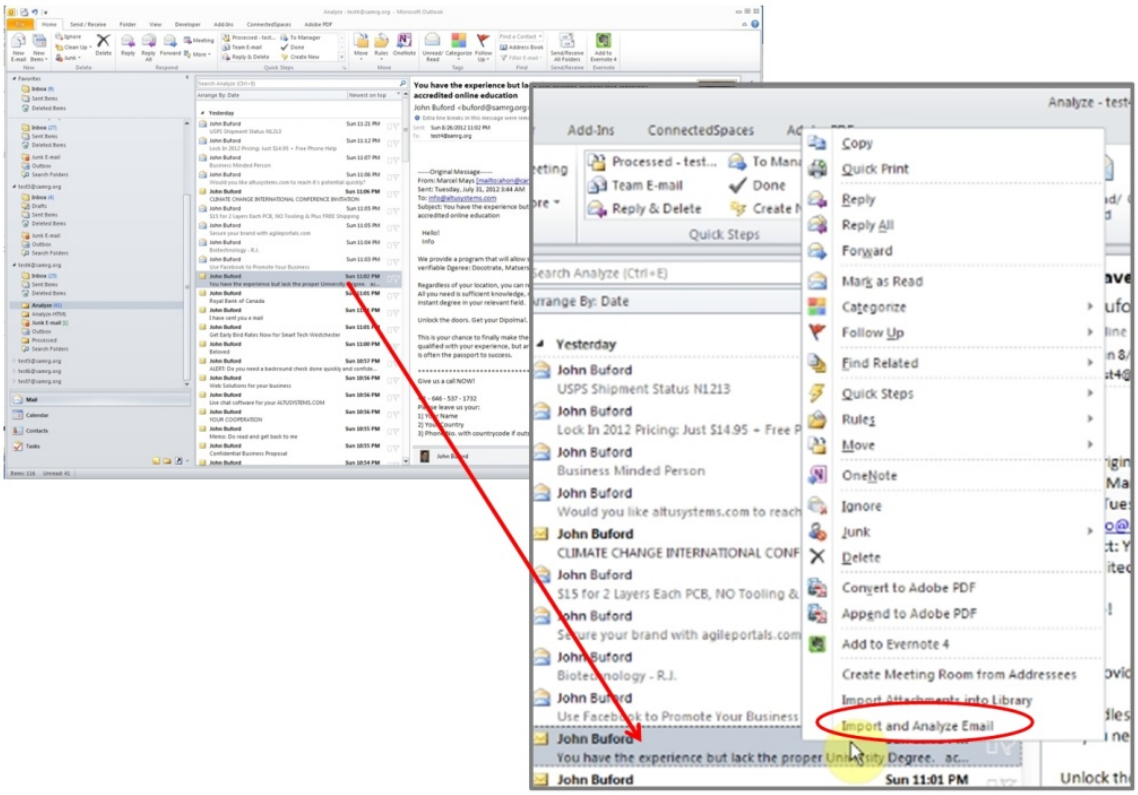


Figure 4 HPPS add-in to Microsoft Outlook 2010.

Figure 5 Email analysis semantic wiki summary page

3.10 Prototype Testing

Approximately 100 emails were processed by the system, representing phishing and non-phishing emails. Representative cases are summarized in the following sub-sections.

Figure 6 Sample Nigerian scam email

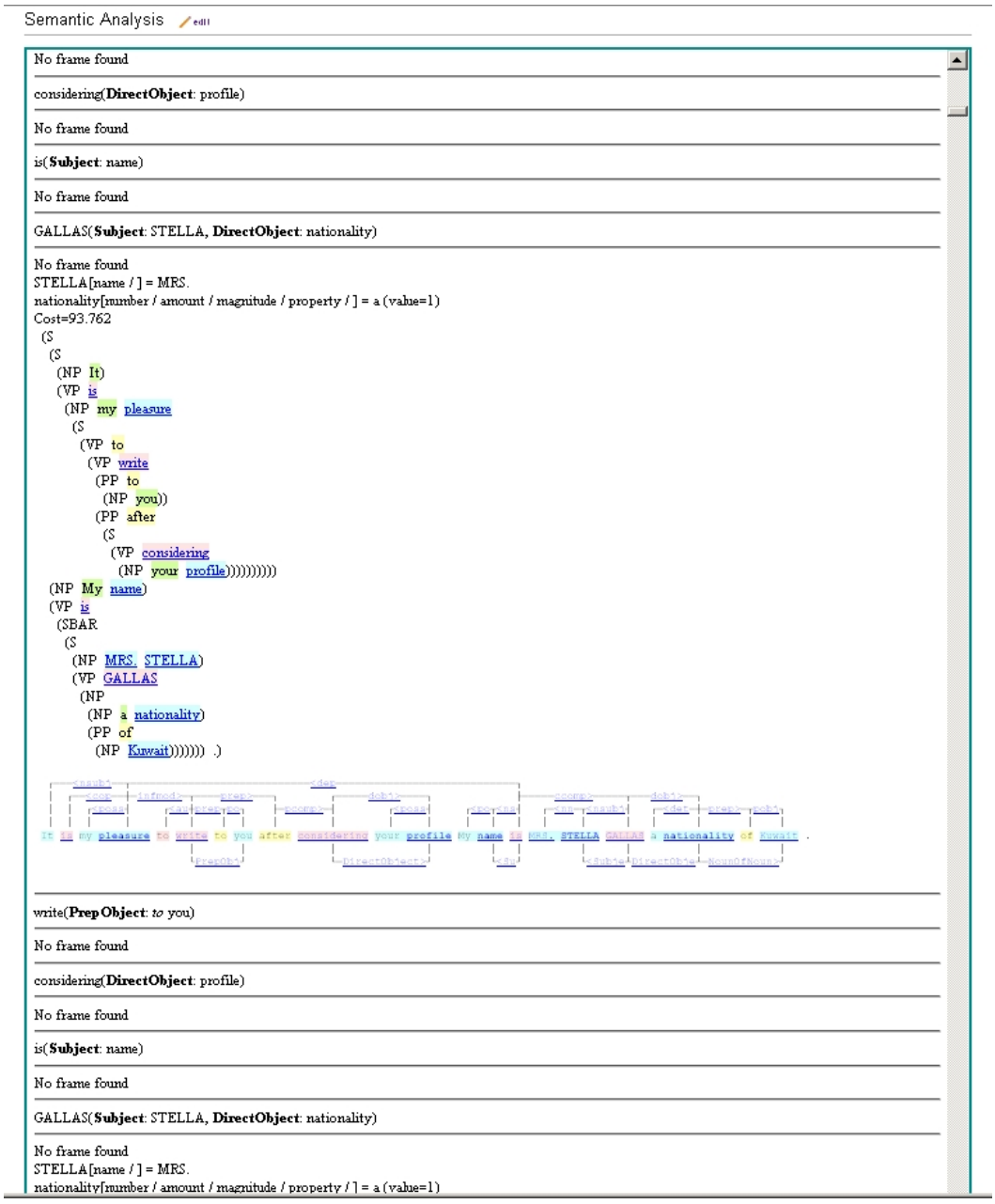


Figure 7 Part 2 – Semantic analysis of the email

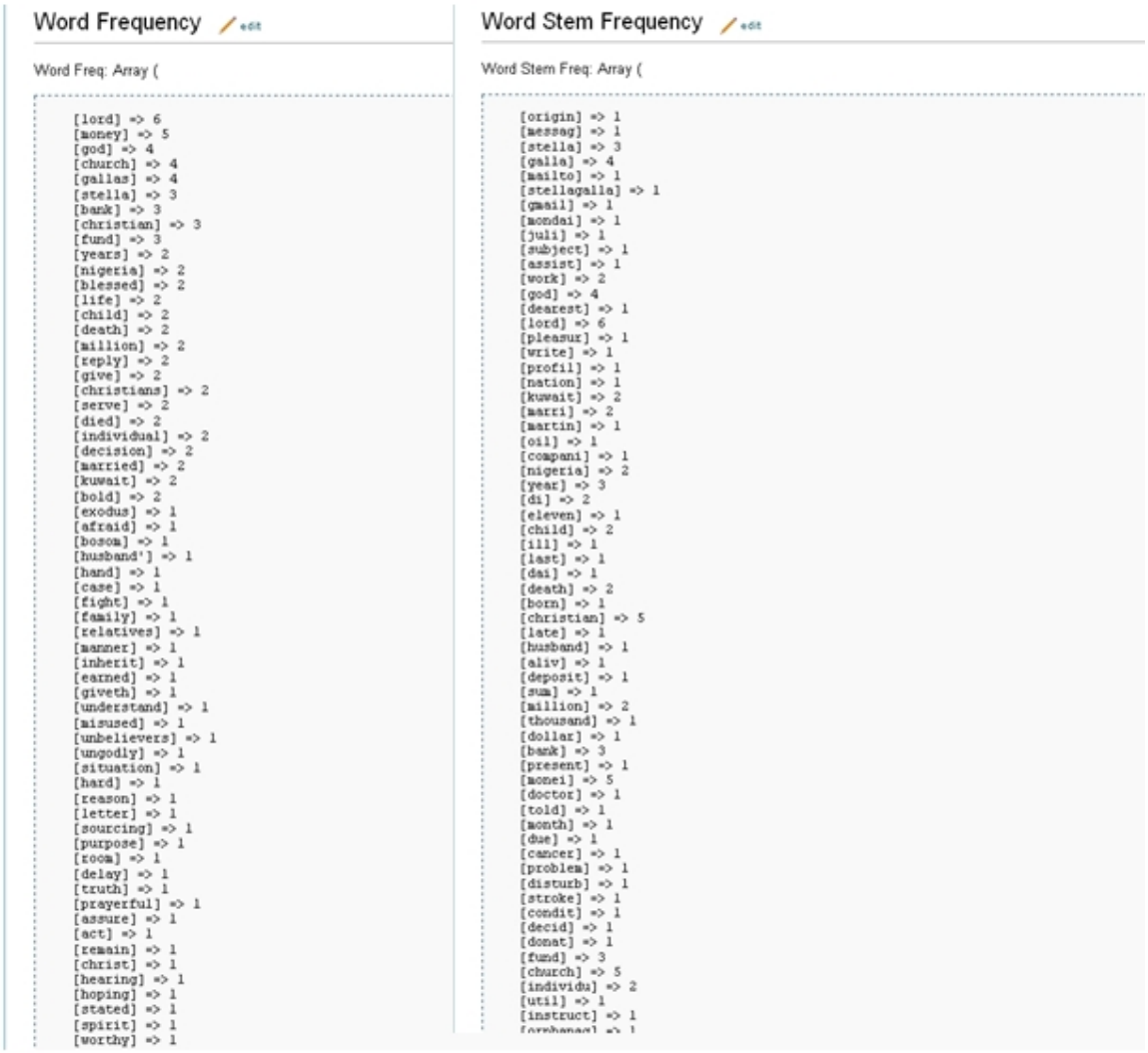


Figure 8 Part 3 – word frequency analysis

Category: Email
Facts about Can You Please $John Buford$2012826-23-42-39-2120000RDF feed

Authority	king +, doctor +, prof +, usa + and bank +
AuthorityCount	7 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	Can You Please Assist Me To Work For God? +
MisspelledWords	gallas stellagallas gmail gallas gallas giveth gallas +
MostFreqStemWords	origin +, messag +, stella +, galla +, mailto +, stellagalla +, gmail +, mondai +, juli + and subject +
MostFreqWords	lord +, money +, god +, church +, gallas +, stella +, bank +, christian +, fund + and years +
NumberMisspelledWords	7 +
Recipient	test4@samrg.org +
Salutation	dearest +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
Urgency	delay +, late +, act +, death + and illness +
UrgencyCount	6 +
WordCount	174 +

Figure 9 Part 4 – summary of properties of the email

3.10.1 **Non-Phishing Marketing**

Email:

-----Original Message-----
From: Katherine Long [mailto:katherinelong80@gmail.com]
Sent: Monday, July 16, 2012 1:13 AM
To: John Buford
Subject: A graphic on the truth about piracy

Hi John Buford,

My name is Katherine and I came across p2pna.com after searching for people that have referenced or mentioned issues related to the piracy of music and movie downloading . I am part of a team of designers and researchers that designed a graphic which highlights how the billions of dollars that Hollywood claims to lose due to piracy, isn't all that they make it out to be. In fact, it may be helping them.

If this is the correct email and you're interested in using our content, I'd be happy to share it with you. :)

Thank you,

Katherine Long
katherinelong80@gmail.com |

Analysis:

Category: Email
Facts about A graphic on th$John Buford$2012826-23-42-38-4780000RDF feed

Authority	bill +
AuthorityCount	1 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	A graphic on the truth about piracy +
MisspelledWords	katherinelong gmail katherinelong gmail +
MostFreqStemWords	origin +, messag +, katherin +, long +, mailto +, katherinelong +, gmail +, mondai +, juli + and john +
MostFreqWords	piracy +, katherine +, john +, buford +, ' +, graphic +, gmail +, long +, katherinelong + and lose +
NumberMisspelledWords	4 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	57 +

Email:

-----Original Message-----
From: fastservice@instant-business.net
[mailto:fastservice@instant-business.net]
Sent: Sunday, July 15, 2012 4:19 PM
To: info@altusystems.com
Subject: SMS text messaging services for altusystems.com

SMS text message mobile marketing for altusystems.com - Everyone is going mobile, are you?

Using our SMS text message services to market to your customers for pennies has several advantages over other mobile options:

97% of all text messages are opened by the recipient!
85% of your customers have a mobile phone! Of those, nearly half are smartphones with Internet access.
Cell phones outnumber computers by a 4 to 1 ratio.
Everyone is turning to their smartphones to search for things when they need them.

Please Click Here <mailto:smart_phones@itimes.com?subject= website info &body= Please type your name, website address and phone number below (All info please). - Smart Phones Optimization -> for more info on our SMS text messaging services for altusystems.com

Smart Phones Optimization

To leave our list go here
<mailto:removeme@instant-business.net?subject=Leave list (Please allow 24 hours)>

3200 Southwest Freeway
Houston, Texas
77027

Analysis:

Category: Email
Facts about SMS text messag$John Buford$2012826-23-42-38-3840000RDF feed

AuthorityCount	0 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	SMS text messaging services for altusystems.com +
MisspelledWords	sms altusystems removeme +
MostFreqStemWords	type +, websit +, address +, phone +, number +, info +, smart +, optim +, sm + and text +
MostFreqWords	smart +, optimization +, leave +, list +, info +, phones +, business +, instant +, houston + and texas +
NumberMisspelledWords	3 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	33 +

3.10.2 **Nigerian 419 Scam**

Email:

-----Original Message-----
From: Hafid Zulaytini [mailto:hafidzulaytini1@gmail.com]
Sent: Thursday, July 19, 2012 12:31 PM
To: undisclosed-recipients:
Subject: URGENT REQUEST; URGENT ANSWER, PLEASE

From: Dr. Hafid Zulaytini;

Dear Friend,

My name is Dr. Hafid Zulaytini; I am in possession of the sum of
US$175,000,000.00 (ONE HUNDRED AND SEVENTY FIVE MILLION U.S DOLLARS) which I wish to keep under your care until I am ready to re-possess it. I was the formal finance minister in Libyan under late Muammar GaddafI. My family and I are presently in hiding in a country in Africa which I intend to disclose to you later. We are running out of fund and I need a part of this money urgently.
I need your assistance to help be contact the security company where this fund is deposited under an open beneficiary status. My position presently cannot allow me do it myself that I why I need someone whom cannot be linked with me in anyway to do it for me.
I will give you 20% of the total fund I have all the necessary information including the deposit code.

Please get back to me for further details.

Dr. Hafid Zulaytini

Analysis:

Category: Email

Facts about URGENT REQUEST;$John Buford$2012826-23-42-38-8840000RDF feed

Authority	minister + and security company +
AuthorityCount	2 +
Benevolence	friend +
BenevolenceCount	1 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	URGENT REQUEST; URGENT ANSWER, PLEASE +
MisspelledWords	hafid zulaytini hafidzulaytini gmail hafid zulaytini hafid zulaytini muammar hafid zulaytini +
MostFreqStemWords	origin +, messag +, hafid +, zulaytini +, mailto +, hafidzulaytini +, gmail +, thursdai +, juli + and pm +
MostFreqWords	zulaytini +, hafid +, fund +, dr +, urgent +, presently +, part +, money +, assistance + and contact +
NumberMisspelledWords	11 +
Recipient	test4@samrg.org +
Salutation	dear + and friend +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
Urgency	late +, urgent + and running +
UrgencyCount	4 +
WordCount	75 +

Email:

-----Original Message-----
From: Mrs. Cassandra Chandler [mailto:robertmueler@fbi.org]
Sent: Saturday, July 21, 2012 3:36 PM
Subject: From Federal Bureau of Investigation (FBI) Contact John Will

Federal Bureau of Investigation (FBI)
Anti-Terrorist And Monitory Crime Division.
Federal Bureau Of Investigation.
J.Edgar.Hoover Building Washington Dc
Customers Service Hours / Monday To Saturday Office Hours Monday to Saturday:
Dear Beneficiary,
Series of meetings have been held over the past 7 months with the secretary general of the United Nations Organization. This ended 3 days ago. It is obvious that you have not received your fund which is to the tune of $2,500,000.00 due to past corrupt governmental Officials who almost held the fund to themselves for their selfish reason and some individuals who have taken advantage of your fund all in an attempt to swindle your fund which has led to so many losses from your end and unnecessary delay in the receipt of your fund.
The National Central Bureau of Interpol enhanced by the United Nations and Federal Bureau of Investigation have successfully passed a mandate to the current president of Nigeria his Excellency President Good luck Jonathan to boost the exercise of clearing all foreign debts owed to you and other individuals and organizations who have been found not to have receive their Contract Sum, Lottery/Gambling, Inheritance and the likes. Now how would you like to receive your payment? Because we have two method of payment which is by Check or by ATM card?
ATM Card: We will be issuing you a custom pin based ATM card which you will use to withdraw up to $3,000 per day from any ATM machine that has the Master Card Logo on it and the card have to be renewed in 4 years time which is 2016.
Also
with the ATM card you will be able to transfer your funds to your local bank account. The ATM card comes with a handbook or manual to enlighten you about how to use it. Even if you do not have a bank account.
Check: To be deposited in your bank for it to be cleared within three working days. Your payment would be sent to you via any of your preferred option and would be mailed to you via FedEx. Because we have signed a contract with FedEx which should expire by August 8th 2012 you will only need to pay $155 instead of
$440
saving you $285 So if you pay before August 8th 2012 you save $285 Take note that anyone asking you for some kind of money above the usual fee is definitely a fraudsters and you will have to stop communication with every other person if you have been in contact with any. Also remember that all you will ever have to spend is $155.00 nothing more! Nothing less! And we guarantee the receipt of your fund to be successfully delivered to you within the next 24hrs after the receipt of payment has been confirmed.
Below are few list of tracking numbers you can track from FedEx website to confirm people like you who have received their payment successfully.
Name: GARCIA .E: FEDEX Tracking Number: 875785927180 (www.fedex.com)
Name: BELINDA DAVIS:FEDEX T racking Number: 876555810411(www.fedex.com)
Note: Everything has been taken care of by the Federal Government of Nigeria, The United Nation and also the FBI and including taxes, custom paper and clearance duty so all you will ever need to pay is $155.
DO NOT SEND MONEY TO ANYONE UNTIL YOU READ THIS: The actual fees for shippingyour ATM card is $440 but because FedEx have temporarily discontinued the C.O.D which gives you the chance to pay when package is delivered

for international shipping We had to sign contract with them for bulk shipping which makes the fees reduce from the actual fee of $440 to $155 nothing more and no hidden fees of any sort!
To effect the release of your fund valued at $2,500,000.00 you are advised to contact our correspondent in Africa the delivery officer Mr. JOHN WILL with the information below, MR JOHN WILL
Email: johnwill77@yahoo.cn
Cell Phone: +234 807 496 8593
You are advised to contact him with the information's as stated below:
Your full Name.....
Your Address:.....
Home/Cell Phone:.....
Preferred Payment Method (ATM / Cashier Check) Upon receipt of payment the delivery officer will ensure that your package is sent within 24 working hours. Because we are so sure of everything we are giving you a 100% money back guarantee.
Yours Sincerely,
Mrs. Cassandra Chandler
FEDERAL BUREAU OF INVESTIGATION
UNITED STATES DEPARTMENT OF JUSTICE
WASHINGTON, D.C. 20535
Note: Do disregard any email you get from any impostors or offices claiming to be in possession of your ATM CARD, you are hereby advice only to be in contact with Mr JOHN WILL of the ATM CARD CENTRE who is the rightful person to deal with in regards to your ATM CARD PAYMENT and forward any emails you get from impostors to this office so we could act upon and commence investigation.

Analysis:

Category: Email
Facts about From Federal Bu$John Buford$2012826-23-42-39-1180000RDF feed

Authority	federal government +, federal +, fbi +, federal bureau of investigation +, account +, tax +, president +, general +, king +, officer +, official +, officials +, secretary +, united nations +, justice +, bank +, yahoo +, fedex +, guarantee +, government +, division +, clearance +, contract + and lottery +
AuthorityCount	58 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	From Federal Bureau of Investigation (FBI) Contact John Will +
MisspelledWords	robertmueler shippingyour johnwill +
MostFreqStemWords	origin +, messag +, cassandra +, chandler +, mailto +, robertmuel +, fbi +, org +, saturdai + and juli +
MostFreqWords	atm +, card +, fedex +, payment +, fund +, federal +, bureau +, investigation +, contact + and receipt +
NumberMisspelledWords	3 +
Recipient	test4@samrg.org +
Salutation	dear +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
Urgency	delay +, expire +, chance + and act +
UrgencyCount	4 +
WordCount	349 +

3.10.3 **College Degree**

Email:

-----Original Message-----
From: Marcel Mays [mailto:ahon@carefirstseniors.com]
Sent: Tuesday, July 31, 2012 3:44 AM
To: info@altusystems.com
Subject: You have the experience but lack the proper University Degree.
accredited online education

Hello!
Info

We provide a program that will allow someone with sufficient work experience to obtain a fully verifiable Dgeree: Docotrate, Matsers or Bacehlors.

Regardless of your location, you can receive a degree in your desired field.
All you need is sufficient knowledge, military, or professional experience and you are on your way to an instant degree in your relevant field.

Unlock the doors. Get your Dipolma!. No time wasted!.

This is your chance to finally make the right move and receive your due benefits. If you are more than qualified with your experience, but are lacking that prestigious piece of paper known as a diploma that is often the passport to success.

****
Give us a call NOW!

+ 1 - 646 - 537 - 1732
Please leave us your:
1) Your Name
2) Your Country
3) Phone No. with countrycode if outside USA

We will get back to you ASAP
****

Do Not Reply to this Email.
We do not reply to text inquiries, and our server will reject all response traffic.
We apologize for any inconvenience this may have caused you.

Analysis:

Category: Email
Facts about You have the ex$John Buford$2012826-23-12-19-1420000RDF feed

Authority	irs +, king +, prof +, usa + and senior +
AuthorityCount	5 +
Benevolence	apologize +
BenevolenceCount	1 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	You have the experience but lack the proper University Degree. accredited online education +
MisspelledWords	ahon carefirstseniors altusystems dgeree docotrate matsers bacehlors dipolma countrycode usa +
MostFreqStemWords	origin +, messag +, marcel +, mai +, mailto +, ahon +, carefirstsenior +, tuesdai +, juli + and info +
MostFreqWords	experience +, degree +, info +, receive +, sufficient +, field +, reply +, prestigious +, piece + and success +
NumberMisspelledWords	10 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
Urgency	chance + and asap +
UrgencyCount	2 +
WordCount	87 +

3.10.4 **System Administration**

Email purporting to be from system administrator:

From: Silvia Reichenback
Sent: Wednesday, August 22, 2012 2:01 AM
To: Silvia Reichenback
Subject:
Your mailbox has exceeded the storage space is determined by the administrator, and you will not be able to receive new messages because we are upgrading from oul2000hn to the new oul5602hn due to some third party trespassing you recognize valid. Tot Re-Validate -> Click here or your account will be block in the less 24hrs.
Thank you. Help Desk

Analysis:

Category: Email
Facts about No Subject$John Buford$2012827-00-10-08-0170000RDF feed

Authority	account +, administrator + and admin +
AuthorityCount	3 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	27 August 2012 +
HasSubject	No Subject +
MisspelledWords	reichenback reichenback hn hn caspio appkey +
MostFreqStemWords	silvia +, reichenback +, wednesdai +, august +, subject +, mailbox +, exceed +, storag +, space + and determin +
MostFreqWords	oul +, hn +, silvia +, reichenback +, http +, caspio +, tot +, dp +, hyperlink + and validate +
NumberMisspelledWords	6 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	40 +

Email purporting to be from system administrator

-----Original Message-----
From: TECH SUPPORT TEAM [mailto:rgunday@omu.edu.tr]
Sent: Saturday, July 14, 2012 1:28 PM
To: undisclosed-recipients:
Subject: Important update #HG89J

Dear email user,

We are undergoing over-congestion due to the anonymous registration of email accounts so that we close some accounts and your account was among those to be deleted.We send you this email so you can verify and let us know if you are currently using this account. To confirm click My account <https://docs.google.com/spreadsheet/viewform?formkey=dGsxSDhYQ1lZdTVhR05MZ1JlZkJNb0E6MQ#gid=0> and submit your credentials and click confirm usage.

Due to the congestion in our webmail servers we are removing all unused accounts, Our webmail administrative team will be shutting down all unused accounts, you must confirm your e-mail account within 72 hours for security reasons. Sorry for the inconvenience this might cost you.

Sincerely,
Email administrator.
powered by Google Copyright 2012©.

Analysis:

Category: Email
Facts about Important updat$John Buford$2012826-23-42-38-2280000RDF feed

Authority	account +, administrator +, admin +, usa + and google +
AuthorityCount	14 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	Important update #HG89J +
MisspelledWords	rgunday omu edu tr viewform formkey dgsxsdhyq lzdtvhr mz jlzkjnb webmail webmail +
MostFreqStemWords	origin +, messag +, tech +, support +, team +, mailto +, rgundai +, omu +, tr + and saturdai +
MostFreqWords	accounts +, account +, email +, confirm +, due +, congestion +, click +, team +, google + and unused +
NumberMisspelledWords	12 +
Recipient	test4@samrg.org +
Salutation	dear +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
Urgency	important +
UrgencyCount	1 +
WordCount	83 +

3.10.5 **Short Phishing Emails**

Email:

> -----Original Message-----
> From: alaa_alqat@yahoo.com [mailto:alaa_alqat@yahoo.com]
> Sent: Tuesday, July 31, 2012 5:59 PM
> To: info@altusystems.com
> Subject: ALERT: Do you need a backround check done quickly and confidentially?
>
> Is Your Arrest Record Posted Online? <http://2729730089.s3-website-us-west-1.amazonaws.com/?s=200>

Analysis:

Category: Email
Facts about ALERT- Do you n$John Buford$2012826-23-12-19-5640000RDF feed

Authority	amazon +, yahoo + and confidential +
AuthorityCount	4 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	ALERT: Do you need a backround check done quickly and confidentially? +
MisspelledWords	alaa alqat alaa alqat altusystems backround amazonaws +
MostFreqStemWords	origin +, messag +, alaa +, alqat +, yahoo +, mailto +, tuesdai +, juli +, pm + and info +
MostFreqWords	yahoo +, alaa +, alqat +, record +, arrest +, confidentially +, quickly +, posted +, website + and amazonaws +
NumberMisspelledWords	7 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
Urgency	alert + and quickly +
UrgencyCount	2 +
WordCount	28 +

Email:

> -----Original Message-----
> From: lilaznbuterfly@yahoo.com [mailto:lilaznbuterfly@yahoo.com]
> Sent: Friday, August 03, 2012 11:17 AM
> To: inite@grm.net
> Subject: You Have Been Sent an E-Card!
>
> See Your Note Here: http://bfhuv.vvlg.550vu.tk

Analysis:

Category: Email
Facts about You Have Been S$John Buford$2012826-23-12-20-7200000RDF feed

Authority	yahoo +
AuthorityCount	2 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	26 August 2012 +
HasSubject	You Have Been Sent an E-Card! +
MisspelledWords	lilaznbuterfly lilaznbuterfly inite grm bfhuv wlg tk +
MostFreqStemWords	origin +, messag +, lilaznbuterfli +, yahoo +, mailto +, fridai +, august +, init +, grm + and net +
MostFreqWords	yahoo +, lilaznbuterfly +, http +, note +, card +, bfhuv +, vu +, tk +, subject + and wlg +
NumberMisspelledWords	7 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	20 +

3.10.6 **Social Phishing**

Email purporting to be birthday greetings from friends

-----Original Message-----
From: tim522@charter.net [mailto:tim522@charter.net]
Sent: Tuesday, September 14, 2010 12:42 AM
To: johnbuford@gmail.com
Subject: Fwd: Happy Birthday John
> From: "Carol Lockhart" <carollockhart@windstream.net>
> To: <tim522@charter.net>
> Subject: Happy Birthday John
> Date: Thu, 2 Sep 2010 06:28:57 -0500
>
> Tell John Happy Birthday and have some fun.
>
> Love,
>
> Carol, Ronny, Roxie &Stephy
>

Analysis:

Category: Email
Facts about Happy Birthday $John Buford$2012827-14-25-17-6180000RDF feed

AuthorityCount	0 +
BenevolenceCount	0 +
Consequences	lock +
ConsequencesCount	2 +
Dataset	Spam +
HasCreationDate	27 August 2012 +
HasSubject	Happy Birthday John +
MisspelledWords	johnbuford gmail lockhart carollockhart windstream sep stephy +
MostFreqStemWords	origin +, messag +, tim +, charter +, net +, mailto +, tuesdai +, septemb +, johnbuford + and gmail +
MostFreqWords	net +, happy +, john +, birthday +, tim +, charter +, subject +, carol +, sep + and thu +
NumberMisspelledWords	7 +
Recipient	test4@samrg.org +
Sender	buford@samrg.org +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	42 +

Email purporting to be from colleague:

-----Original Message-----

From: Khushboo Bohacek [mailto:khushboo000@gmail.com]
Sent: Tuesday, July 05, 2011 2:52 PM
To: 3-Sixty Movers
Subject: This is embarrassing. Anyway please ignore.

Please ignore the email below.

Thanks,
Khushboo

Dear Friends:
I have good news for you. Last week.
I have orders China Quantity: 21 Products Apple MacBook Pro MB986LL / A I received the Apple MacBook Pro MB986LL / A Product!
web: www.gaoshujing.com
It's amazing! The article is original, new, and has high quality t, but it is muc cheaper.
I am pleased with this good news to share with you!
Sincere!

Analysis:

Category: Email
Facts about FW- This is emb$John Buford$2012823-01-49-31-9540000RDF feed

AuthorityCount	0 +
Benevolence	friend +
BenevolenceCount	1 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	23 August 2012 +
HasSubject	FW: This is embarrassing. Anyway please ignore. +
MisspelledWords	khushboo bohacek khushboo gmail khushboo macbook ll macbook ll gaoshujing +
MostFreqStemWords	origin +, messag +, khushboo +, bohacek +, mailto +, gmail +, tuesdai +, juli +, pm + and sixti +
MostFreqWords	khushboo +, news +, good +, ignore +, apple +, ll +, mb +, macbook +, pro + and original +
NumberMisspelledWords	10 +
Recipient	test4@samrg.org +
Salutation	dear +
Sender	buford@altusystems.com +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	54 +

Email purporting to be from colleague with technically appropriate subject matter

From: LundyLewis [mailto:lundylewis142@yahoo.com]
Sent: Tuesday, June 19, 2012 10:32 AM
To: buford@altusystems.com
Subject: Network Security Assessment Of The National

Dear,
Please find attached and give some advice.
http://economic.ned-news.org/Network_Security_Assessment_Of_The_National.zip
Regards,

Analysis:

Category: Email
Facts about FW- Network Sec$John Buford$2012823-01-49-32-3140000RDF feed

Authority	yahoo +
AuthorityCount	1 +
BenevolenceCount	0 +
ConsequencesCount	0 +
Dataset	Spam +
HasCreationDate	23 August 2012 +
HasSubject	FW: Network Security Assessment Of The National +
MisspelledWords	lundylewis lundylewis altusystems +
MostFreqStemWords	lundylewi +, mailto +, yahoo +, tuesdai +, june +, buford +, altusystem +, subject +, network + and secur +
MostFreqWords	security +, assessment +, lundylewis +, network +, national +, http +, advice +, give +, economic + and ned +
NumberMisspelledWords	3 +
Recipient	test4@samrg.org +
Salutation	dear +
Sender	buford@altusystems.com +
SubjunctiveCount	0 +
UrgencyCount	0 +
WordCount	28 +

3.11 Analysis

The prototype system implements most of the recommended heuristics discussed earlier in the report. Some heuristics (urgency, salutation) are seen in testing to be useful discriminators for phishing emails. Other heuristics (misspelling, consequences, benevolence) were not as strong indicators in the particular emails tested. Subjective mood constructions were not found in the test data.

Further improvement to the heuristics is needed for:

- Short emails
- A social model of each user which can be used to predict if specific content from a "friend" is likely, for example using previous email exchanges with that "friend"
- System admin phishing, for example by validating the technical jargon and the source of the email

The prototype can be furthered enhanced and tuned by improving the implementation of the heuristics and by more extensive testing.

4. Meetings with Sponsor

4.1 Kickoff Meeting

The project kickoff meeting was held remotely by teleconference on February 1, 2011. Attending from Altusys were: John Buford and Nina Kohli-Laven.

4.2 Final Meeting

The final project meeting is TBD.

5. Cost Status

The project is within budget.

6. Intellectual Property Developed

Altusys has not filed a patent application at this time.

7. Plan for the Phase II

The goal of Phase II is to extend the models with additional testing and to develop a fully functional implementation.

8. Conclusions

Detailed analysis of phishing and non-phishing email data sets was performed to determine socio-linguistic indicators for phishing emails. A set of heuristics was developed.

The prototype system implements most of the recommended heuristics. Some heuristics (urgency, salutation) are seen in testing to be useful discriminators for phishing emails. Other heuristics (misspelling, consequences, benevolence) were not as strong indicators in the particular emails tested. Subjective mood constructions were not found in the test data.

Further improvement to the heuristics is needed for:

- Short emails
- A social model of each user which can be used to predict if specific content from a "friend" is likely, for example using previous email exchanges with that "friend"
- System admin phishing, for example by validating the technical jargon and the source of the email

The prototype can be furthered enhanced and tuned by improving the implementation of the heuristics and by more extensive testing.

During phase 1, Altusys developed the foundation for key components of the proposed Hybrid Phishing Protection System, focusing on socio-linguistic heuristics that address social-based phishing attacks. Further, HPPS requirements and preliminary architecture were developed as a basis for further research.

9. Bibliography

[1] Stanford Natural Language Processing Group. The Stanford Parser: A statistical parser. Retrieved from: http://nlp.stanford.edu/software/lex-parser.shtml

[2] Martin Porter. Porter Stemming Algorithm. Retrieved from: http://tartarus.org/martin/PorterStemmer/index.html

[3] Jonas Sjöbergh and Kenji Araki, *"Extraction based summarization using a shortest path algorithm"* , 12th Annual Language Processing Conference NLP2006, Yokohama, Japan, 2006.

[4] Hercules Dalianis, Jonas Sjöbergh, and Eriks Sneiders, "Comparing Manual Text Patterns and Machine Learning for Classification of E-Mails for Classification of E-Mails for Automatic Answering by a Government Agency", CICLing 2011

[5] Proxem. Antelope for .Net. Version 0.8.7. March 2009. Retrieved from: http://www.proxem.com/download/Proxem.Antelope.0.8.pdf

[6] WordNet. A Lexical Database for English. Retrieved from: http://wordnet.princeton.edu/

[7] D. Temperley, D. Sleator, J. Lafferty. Link Grammar. Retrieved from: http://www.link.cs.cmu.edu/link/

[8] Enron Email Dataset. Retrieved from: http://www.cs.cmu.edu/~enron/

[9] Semantic Media Wiki. Semantic-mediawiki.org.

[10] SMW+. www.smwplus.com

[11] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining Email Content for Author Identification Forensics. SIMOD Record 30(4): 55-64 (2001);

[12] J. Li, R. Zheng, and H. Chen. From Fingerprint to Writeprint: Feature Selection for Authorship Identification. Communication of ACM, 49(4): 76-82 (2006)

10. Research Team

- **Nina Kohli-Laven (Social Anthropologist), Research Scientist**

  Dr. Kohli-Laven has extensive expertise in designing local data collection, analysis, and training software for the US military on DoD-funded projects. She holds a Ph.D. in Social Anthropology from the University of Michigan, Ann Arbor, and a Master of International Affairs from Columbia University, and trained in socio-linguistics and linguistic anthropology while at Michigan in the Linguistic Anthropology Program at the Department of Anthropology. Her post-doctoral training was at the School of Medicine at McGill University. She speaks and has designed and conducted survey tools with native populations in Arabic, Persian-Dari, Tajiki, French, and Albanian in the Middle East, South Asia, and Balkans. Dr. Kohli-Laven worked on the design of automated social data collection software for military use in Afghanistan with Soar Technologies, the development of cultural instruction and training software for platoon-level military officers in Iraq with Vcom3D Inc. She also led the creation of customized behavioral data collection and analysis processes for adaptable cognitive models of local populations at Soar. She has conducted research with military serving in Iraq and Afghanistan and native Afghans as part of the development of these projects. In 2002, she participated in delivery of a survey methodology and program monitoring tool for needs assessment in post-conflict zones on a joint project of the International Rescue Committee and World Health Organization. The tool is used in humanitarian relief contexts to design local health programs worldwide. Dr. Kohli-Laven conducted field research for her dissertation in Sana'a, Yemen and Dushanbe, Tajikistan as a National Science Foundation Graduate Research Fellow and was a U.S. Fulbright Fellow in 2007-2008. She is also currently senior advisor on field methods and field data analysis to a 5-year cross-disciplinary study of health decision-making funded by the National Institutes of Health Transformative Research Projects funds at the University of California, San Francisco.

- **John Buford, PhD, President**

Dr. Buford is a principal of Altusys. More information can be found at www.altusystems.com

11. Appendices

11.1 Socio-Linguistic Anti-Phishing Feasibility Study Summary

Inquiry A: Is it possible to use socio-linguistic features of phishing emails to generate a phishing "signature" that can be used to alert users when an email is socio-linguistically suspect?

Dataset: 8 known phishing emails from 2011 and 2010 downloaded from the internet – Chase, Wells Fargo, UPS, and Amazon (need more).

Method: Analyze a sample set of known phishing emails for a cross-section of institutions and request types (see Table 1 and 2) looking for patterned structural, stylistic, semantic, and syntactical patterns that may distinguish the intentions of the authors. The task here is to build on John Austin's *speech act theory* and applied work on *linguistic register* to determine if there is an identifiable *phishing email register* and set of *phishing speech acts* (i.e. language that is being used by the author to create certain emotional effects in the world of the reader) that prevail across phishing emails or subsets of phishing emails. Could any patterns in phishing language be used to facilitate detection?

Conclusions: The linguistic features of phishing emails identified in the proposal and kick-off appear valid in this analysis of a different set of emails. See Table 6 for the list of preliminary linguistic features. There are also basic structural and stylistic patterns present:

- Stylistic features: misspellings, numerous repetitions of words, frequent hyperbolic punctuation and language ("!"). In the case of Finance emails, there is much more use of personal pronouns ("we") than legitimate emails from financial institutions.
- Structural features: emails are lengthier and repetitive compared to legitimate communication.

Table 6 Preliminary Linguistic Features Identified and Validated in Months 1 and 2

	Feature	Description	Associated Attribute
1	**Convey Urgency Feature**	Phisher makes the message seem urgent	"immediately," "alert," "as soon as possible"
2	**Communicate Adverse or Advantageous Consequences Feature**	Phisher makes the message seem consequential	"if you do not respond," "please be advised," "account closure," "might result in the"
3	**Assert Authority Feature**	Phisher makes the message seem authoritative	"on file," "our records indicate," "your account at our institution," and other uses of "our" or references to data about the reader that the author owns or

			controls
4	**Communicate Author Benevolence Feature**	Phisher makes the message seem in the reader's best interests	"for your protection" "bring to your attention" "provide you the opportunity to" "protect you"

Inquiry B: Is it possible to build an email signature for individual authors so that an automated email filter could recognize when an email from the individual's email address is consistent with his/her style/signature or not (i.e. distinguish it from phishing)?

Dataset: Enron Email Corpus

Method: Analyzed a sample set of emails for 8 different authors (about 50 each) looking for structural, stylistic, semantic, and syntactical patterns that may distinguish or characterize their unique "speech" style, or "idiolect." Email files for authors were selected, each containing about 400 emails sent in 2000 and 2001 and qualitatively assessed then systematically compared to determine whether semantic and syntactic style was consistent across emails from a single author. Every 5th email in a log of all of the emails was selected for review until 30-40 emails had been reviewed. In cases where emails were short and therefore not empirically rich, review continued until 60-80 emails had been reviewed.

Conclusions: Preliminary analysis of features of sets of emails from the Enron corpus suggest that individual authors use consistent stylistic and structural features:

1. *Yes, Semantic/pragmatic patterns may be able to distinguish speech style between authors and between classes of authors.* The data also suggests that author-specific semantic/pragmatic signatures (e.g. the Harry signature or the Joe signature) could be generalized to classes of authors (e.g. gender appears to be a significant correlate of semantic and pragmatic style). Further investigation of semantics should place the emails in the context of who the recipient is and what the relationship between recipient and sender is (e.g. is professional hierarchy a factor in the Polite Modal for both men and women senders?).

2. *However, it was also evident that, although semantic differences provided cues to author identity, detection of simple linguistic patterns would suffice as a differentiator of authors in all of these cases.* Examples of simpler patterns are lexical choices, as well as format, length, signature, punctuation, and basic syntactical choices (e.g. full sentences with pronouns, prepositions, verbs and nouns vs. sentences that cut-off personal pronouns etc.). Focusing on these simpler linguistic features instead of more complex semantic or pragmatic patterns is also attractive because even short emails can be rich with data about these simpler patterns (semantic and pragmatic analysis is hard to do with 1-line or 2-line samples, and many of the emails in the corpus, and more generally, are quite short.).

Research Recommendation: Focus on linguistic analysis of lexical choices, format, length, signature, punctuation, and basic syntactical choices to build models of individual authorship of emails. This should be sufficient as a differentiator of authors and is usable in both short and long emails (i.e. wide range of applicability to various email lengths).

Over all, this approach also optimizes the use of socio-linguistic tools for email identification problems: overcoming limitations of the bag-of-words method by adding numerous new and inter-related socio-linguistic features to the analysis of email; but, saving needless time and energy that might be invested in a semantic or pragmatic analysis, where added value is unclear. This approach conforms to the new approaches suggested in the literature on linguistics and author identification by del Vel and Zhang.[2]

[2] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining Email Content for Author Identification Forensics. SIMOD Record 30(4): 55-64 (2001); J. Li, R. Zheng, and H. Chen. From Fingerprint to Writeprint: Feature Selection for Authorship Identification. Communication of ACM, 49(4): 76-82 (2006)

11.2 Socio-Linguistic Detection Features of Phishing Emails

Phishing Emails that Pose as Official Communication

The emails usually pose as official communication from a financial or online payments or retail company or institution and dupe recipients into visiting and providing personal financial information.

Principal Features:

Consequences: These features capture how the Phisher makes the message seem consequential. The Phisher does this in two ways.

First: by using subjunctive constructions (possibility subjunctive, purpose subjunctive) (e.g. "if you do not do x, you will…").

Examples:

"if you do not respond"

"if not _____ed immediately"

"if we do not receive _____ by/within, then"

"might result in the"

"please be advised"

Second: by listing consequences of not acting (e.g. account closure, deletion, etc..). These consequences are often expressed hyperbolically (i.e. overstated) through the use of hyperbolic punctuation (e.g. !), definitive negative constructions (e.g. you will not, you cannot), and specified extreme consequences (e.g. deletion, denial, suspension, closure).

Examples:

"account closure"

"will be [verb]ed" (e.g. will be deactivated, will be deleted)

"open an investigation"

"re-activate"

"expiration"

"prevented access"

"suspended"

Failure to do this within [time] will lead to [consequence noun, e.g. suspension/denial/closure]

"locked" "lock"

"stop"

Legitimate emails do not use hyperbolic words, use open constructions (e.g. may not), and non-specific consequence words (impacts, issues, problems).

Urgency: The Phisher uses language to make the message and any response to it urgent and imperative. The phisher makes response to the message seem urgent in three ways.

First: by qualifying time with words such as “now” and “immediately.”

Examples:

“ [imperative verb] now” (e.g. “respond now” “go now” “click now”)

“immediately”

Second: by using constructions that imply time sensitivity such as “Alert” and “Account at risk.”

Third: the Phisher makes response to the message seem imperative by using imperative verb constructions (e.g. Stop, go, login, click), the deontic modal construction (e.g. “you must”), and other regular verbs expressing necessity (e.g. “need”).

Examples:

“needs to be [verb]”

“you need” “need” “needs”

“obligatory”

“you must” “it must” “we must”

“security alert”

“security check”

“we require you to”

“you are required to follow”

“Immediately” is an urgency word that also often appears in legitimate emails. We omit “immediately” from the Urgency diagnostic process for this reason. “Login” is an imperative that also often appears in legitimate emails. We omit “login” from the urgency diagnostic process for this reason. “Alert” is an urgency word that also sometimes appears in legitimate emails. We consider “Alert” less indicative of Urgency than other listed words.

Errors: Errors in the spelling, order, and agreement of different parts of speech within the email text (syntactic and orthographic errors), when compared to standard and accepted usage within the language medium. Altusys examined phishing and legitimate emails in English and French to determine this feature. The feature was evident in the French email set suggesting that it is a feature of Phishing emails that is not specific to English language emails.

There are two types of error:

Regular error: the error is sometimes regular error in the syntax or orthography of English words. Patterned error is almost always indicative of non-native usage.

Examples:

Non-agreement of subject and verb (agreement according to rules in non-English languages that conflict with English rules)

Misplaced infinitives, e.g. “to prevent this to happen” (many non-English languages use infinitives to express ideas that are not expressed with infinitives in English)

tendency to pluralize English words that are not used in plural form in English, but that are regularly pluralized in non-English languages.

“cooperations” (кооперация = Russian plural)

“informations” (информация = Russian plural)

Irregular error: the error is haphazard, non-patterned, e.g. random misspellings, absence of punctuation, connecting words (with, to), or pronouns. Non-patterned error is just poor usage and is also a characteristic of many phishing emails.

Examples:

Capital letters in mid-sentence or mid-word

Absence of full stops, capital letters at opening of sentence, excess space between words or sentences

Orthographic and syntactic error would never occur in a legitimate email.

Benevolence: The Phisher makes the message seem in the reader's best interests by using phrases implying the virtuosity, concern, or diligence of the sender in protecting the recipient from harm.

Examples:

"for your protection"

"bring to your attention"

"provide you the opportunity to"

"protect you"

“[bank/institution e.g. “Chase”] safeguards your account”

“state-of-the-art technology”

“Valued customer”

“Valued [bank/institution e.g. “Chase”] customer”

Legitimate emails sometimes use benevolence language, including “security,” “protect,” “your protection,” etc. so this word set will be assigned lesser significance as an indicator.

Authority: The Phisher makes the message seem authoritative by using words that imply authority of the sender of the message. This is usually done in two ways.

First: by strategic use of the ‘pointing; pronouns “we,” “us,” and “our,” especially dense use of “we” and “our” in the opening lines of the message text. In linguistics, this is known as person deixis – pointing to specific people in the utterance or text. The author’s choice to point to “we” instead of using passive/impersonal verb constructions to allude

to himself/herself achieves two principal effects: (1) It implies the author is an authoritative and plural actor (i.e. acting in the name of a body, institution, or other organization that is greater than just him or herself), and (2) It establishes a social relationship between sender and recipient, making the communication more compelling to the recipient.

Examples:

"we have reviewed"

"we consider"

"our records indicate"

"your account at our institution,"

Altusys regards "we" in Line 1 of the body of the email as an indicator of Phishing email likelihood. "We" in subsequent lines is a characteristic of Legitimate as well as Phishing emails so is a less distinctive Phishing indicator. In emails in the test set examined by Altusys, 7.5% of Legitimate emails used "we" in Line 1 while 30% of Phishing emails used "we" in Line 1.

Second: by alluding to the official nature of the communication and the empowered status of the sender through language that implies omnipresence and oversight.

Examples:

"on file"

"official notification that…"

Difference with legitimate emails: the legitimate banking emails reviewed did not use "we" in the opening lines of the message text except for one bank, Chase, which regularly uses "we" throughout message texts in routine and alert emails. Online services Skype and Facebook use "we" in opening lines of message text while Yahoo and PayPal do not.

Responsibility: The Phisher increases the likelihood that the recipient will feel it is incumbent on him/her to personally take action by using words that personally implicate the recipient in the message.

The Phisher achieves this by use of the 'pointing; pronouns "you" and "your." As in the case of the Authority feature described above, this is known as person deixis – pointing to specific people in the utterance or text. The author's choice to point to "you" instead of using passive/impersonal verb constructions has two functions: (1) It conveys incumbency on the recipient to *personally* act to address the topic of the email, and (2) It establishes a social relationships between sender and recipient, making the communication more compelling to the recipient. The "you" functions deictically in that it indicates the problem is with YOU, personally. Used in this way, it's a form of persuasion.

Examples:

"You sent a payment"

"This email has been sent to you "

"…the way we serve you…"

Altusys regards "you" in Line 1 of the body of the email as an indicator of Phishing email likelihood. "You" in subsequent lines is a characteristic of Legitimate as well as Phishing emails so is a less distinctive Phishing indicator. In emails in the test set examined by Altusys, 23% of Legitimate emails used "you" in Line 1 while 37% of Phishing emails used "you" in Line 1.[3] The majority of the Legitimate emails that used "you" in Line 1 were marketing emails (e.g. advertisements from banks and online retailers for new services or technologies)(See Assumptions Section 4 for further discussion of marketing emails).

Salutation: Phishing emails initiate the communication differently than Legitimate emails (See Table 7).

Table 7 Salutation style in Phishing and Legitimate emails

Frequency	Legitimate Emails	Phishing Emails
Most common	No Salutation	Dear [bank/institution] client, member, customer, card holder, cardholder, seller, buyer, account holder Dear valued ….
Sometimes	Dear [Full registered name of recipient]	Dear [full email address of recipient]

Tense: Phishing emails sometimes use gerunds and relative tenses (linguistic reason unclear). Official emails rarely use gerunds and almost never use relative tenses. Banks are less likely to use gerunds than online services such as facebook, Amazon, Paypal, eBay etc. (See Section 4. Assumptions, #2). Gerunds are stylistically less formal language.

Examples:

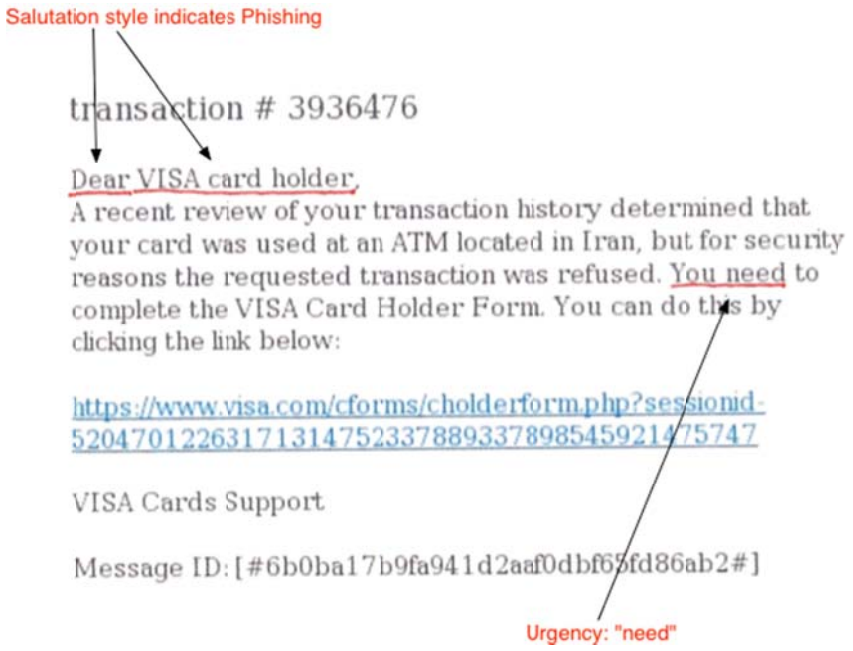
Pluperfect constructions like "you had used your account" Legitimate emails use absolute tenses, e.g. simple present, past, future, "use" "used" "will use."

Gerunds like "responding" "protecting" "having" "traveling"

[3] When two separate Phishing emails in the test set were similar (i.e. same message, structure, and vocabulary with only minor modifications to format or phrasing) we counted the instance of "you" in both examples as 1 instance.

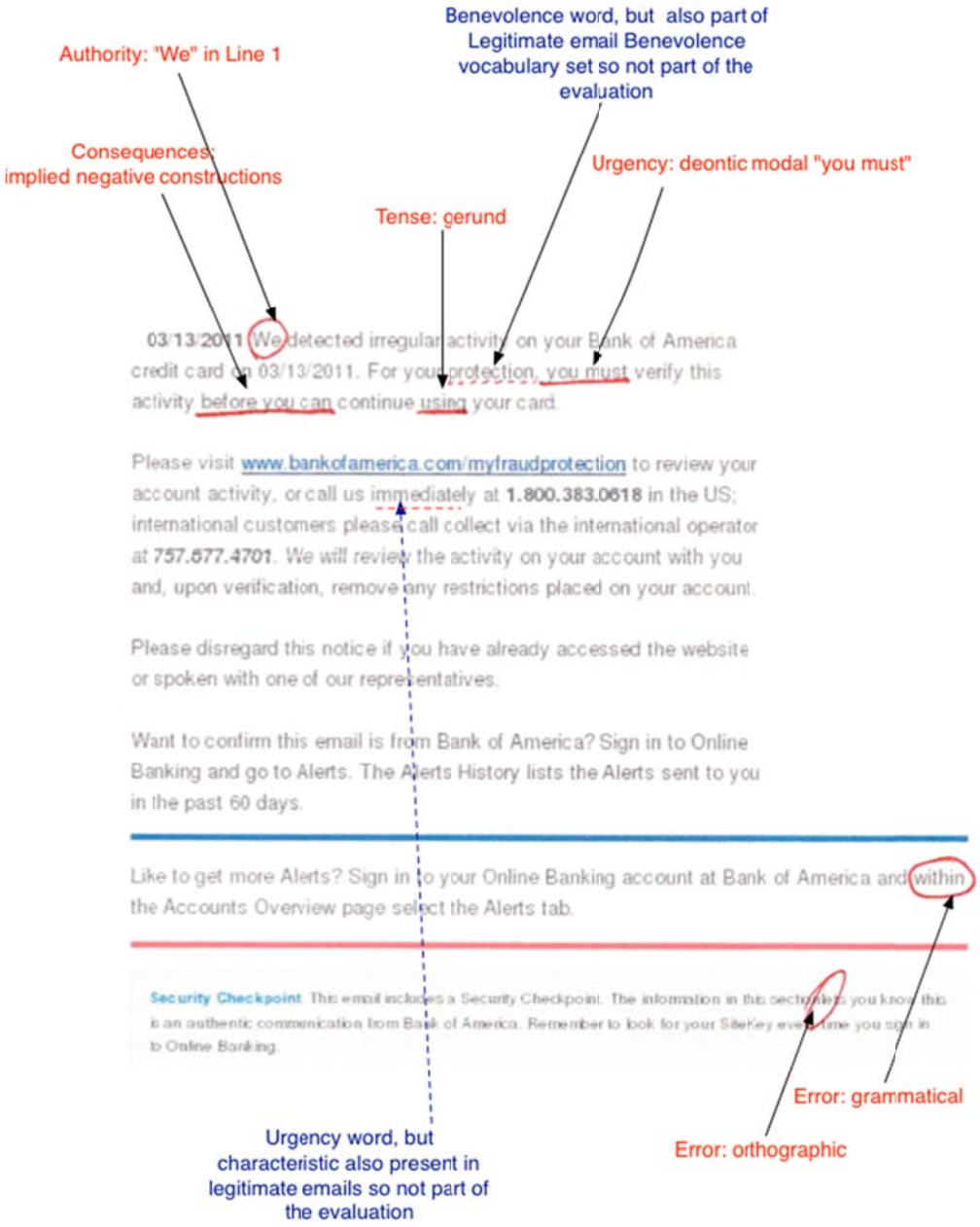
11.3 Diagrammed Selections From the Test Set

11.3.1 **Visa Phishing Email**



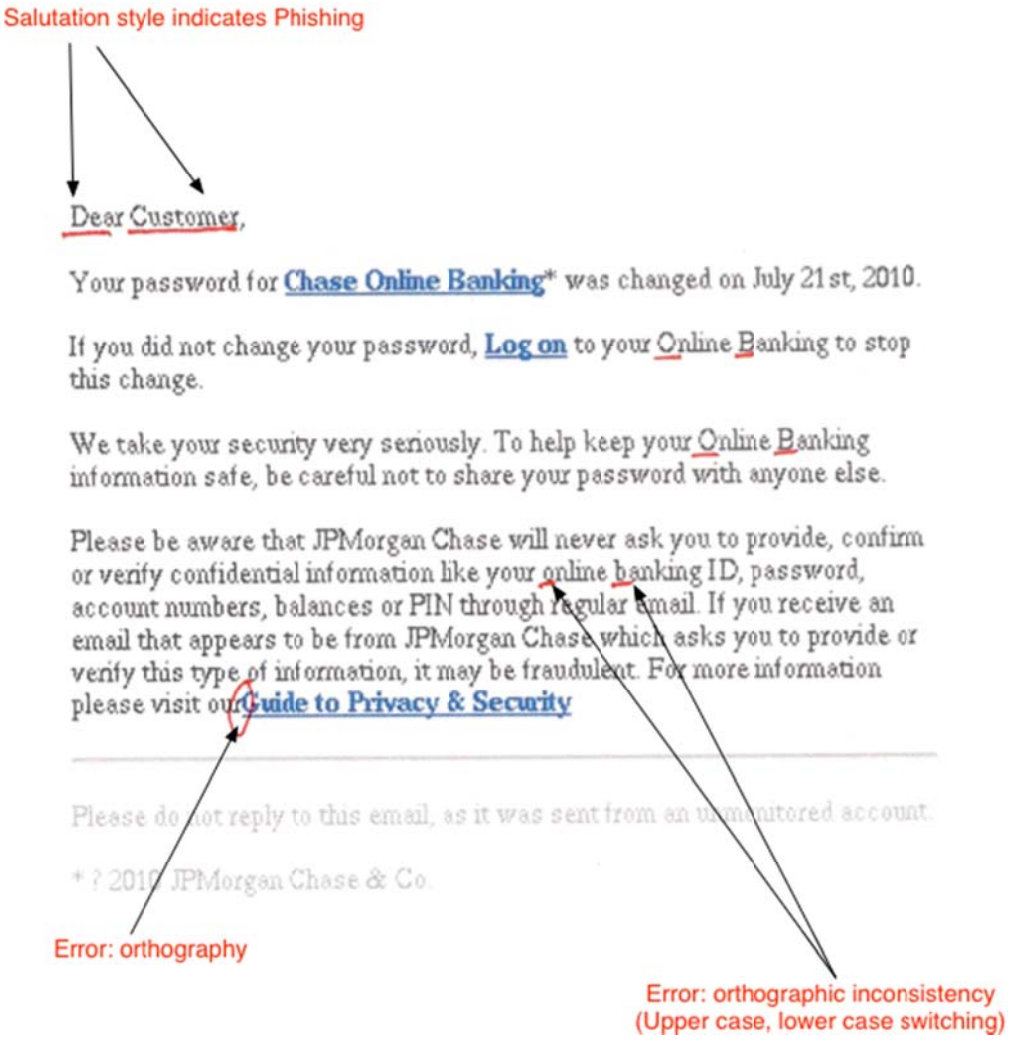
<u>Socio-Linguistic Evaluation</u>:
Salutation Grade 1
Urgency Grade 1

11.3.2 **Bank of America Phishing Email**



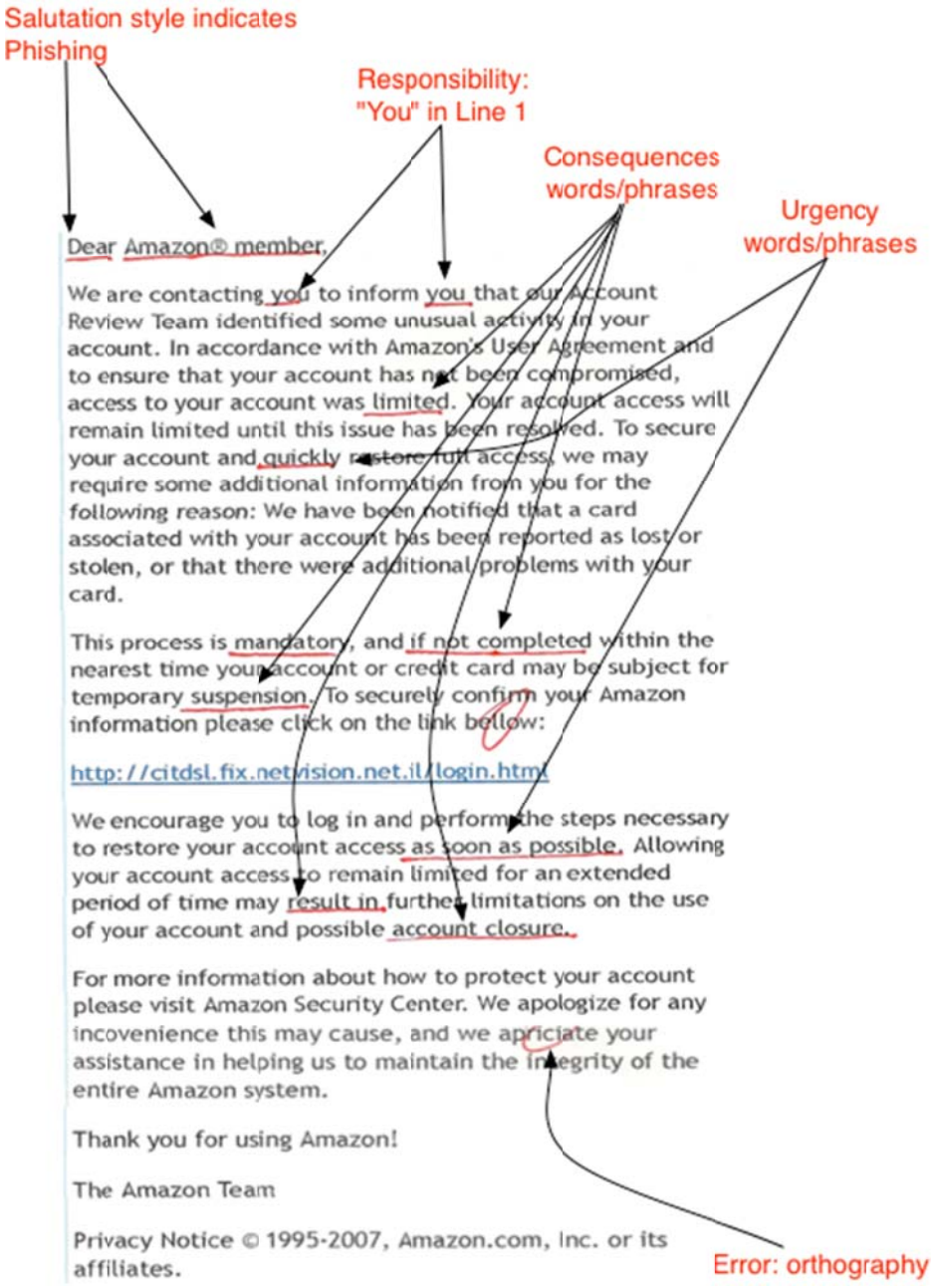
<u>Socio-Linguistic Evaluation</u>:
Consequences Grade 2
Urgency Grade 2
Tense Grade 2
Authority Grade 2
Benevolence Grade 2
Error

11.3.3 **Chase Phishing Email**



Socio-Linguistic Evaluation:
Salutation Grade 1
Error

11.3.4 **Amazon Phishing Email**



<u>Socio-Linguistic Evaluation</u>:
Consequences Grade 1
Salutation Grade 1
Urgency Grade 1
Responsibility Grade 1
Error

11.3.5 **Legitimate Citibank Email**

11.4 Grading the 8 Socio-Linguistic Features of Phishing Emails

Following is a list of words, phrases, word combinations, and grammatical types that, when present, signify Grade 1 (higher) or Grade 2 (lower) risk that an email is Phishing.

Consequences:

Grade 1:

"if" AND "not" OR "do not" OR "is not" OR "then" OR "by"

"suspended" OR "failure" OR "lock" OR "locked" OR "block" OR "blocked" OR "deactivated" OR "deleted" OR "closed" OR "closure" OR "expired" OR "close"

"!!" OR "!!!" OR "!!!" OR "!!!!"

"result in" OR "lead to" OR "cause"

Grade 2:

"!"

"Will be" OR "in order to" OR "will not" OR "before" OR "will never" OR "!"

See flow chart for illustration of how Grades operate in the evaluation of the probability that the email is Phishing (Fig. 2).

Urgency:

Grade 1:

"need" OR "needs" OR "obligatory" OR "requirement" OR "requires" OR "require" OR "required" OR "must" OR "as soon as possible" OR "quickly" OR "right away"

Grade 2:

"now" OR "immediately"

"Alert" AND "now"

"[command verb]" AND "Alert"

Errors:

Grade 1:

Erroneous orthography

Non-agreement of subject-verb

Misplaced infinitives

Benevolence:

Grade 1:

"value" OR "values" OR "valued"

Grade 2:

"provide" OR "provides" OR "protect" OR "protects" OR "protection" OR "safeguard" OR "safeguards"

Authority*:

Grade 1:

"file" OR "official" OR "records"

Grade 2:

"notification" OR "our [noun e.g. bank, company]" OR "we" (Social networking and certain online retailers and services use "we" frequently, while banks use it rarely.)**

Responsibility*:

Grade 1 Phishing:

If the number of uses of "you" in Line 1 and Line 2 exceeds 0

Grade 1 *Not-Phishing*:

Passive verb constructions and past participles (i.e. presence of these types suggests the email is likely to be legitimate, not Phishing. Phishers are more likely to use active constructions that point to the reader than passive constructions).[4]

Salutation:

Grade 1:

"Dear" AND "client" OR "member" OR "customer" OR "card holder" OR "seller" OR "buyer" OR "account holder"

"Dear" AND "valued"

Grade 2:

"Dear" AND [email address]

Tense:

Grade 1:

Relative tenses (pluperfect)

Grade 2:

Gerunds (except in the case of social networking and some online retailers and services, which user gerunds to convey informality)**

Further Notes

*** Catching phishing emails at the expense of preventing some marketing emails from reaching the user**. Although most legitimate emails use "you" and "we" less than phishing emails, there are some legitimate emails – especially marketing emails from

[4] Examples of regular past participles are: *hired, worked, logged, addressed, informed, replied, responded, required, accessed, received, notified, downloaded, signed, changed, updated, contacted, activated, deleted, deactivated, protected, provided, confirmed, verified*. Irregular past participles that are common in legitimate emails are: done, said, written, seen, sent, reset.

banks and other trusted institutions – that use "you" much the same way as phishing emails do (i.e. setting up a personal relationship of responsibility between the sender and recipient and urging action through the use of deictic language and urgent phrases and words). This is not necessarily surprising: Making communication feel personal for the reader and making the reader feel it is incumbent on him/her to personally respond is a marketing tactic just as much as it is a phishing tactic. When designing the Authority and Responsibility indicators and rules, we have provisionally viewed user receipt of marketing emails from legitimate institutions as lower priority than user receipt of other emails from legitimate institutions.

** **Defining institution or institutional-class specific rules for detection of legitimate emails.** Institutions such as banks or online service companies tend to have linguistically consistent, within a certain range, communication across emails over time. This could be described as institution-specific *registers*: ways of writing and addressing members/customers that are a manifestation of the particular communication style, aims, and norms within a particular bank or company. Registers may differ from one institution to another. For instance, Citibank and Chase have slightly different modes of addressing the user/recipient. Registers differ more significantly across different classes of institution. For example, major banks such as Citibank, Bank of America, Chase, etc. have similar registers, over all, to each other but somewhat different registers from online services such as Mint.com, Paypal, or Facebook. While the banks use formal, bureaucratic, impersonal, and passive language constructions, the latter set of companies often use colloquial, informal, personal, and active language constructions. Altusys finds it useful to assume that rules enabling the evaluation of legitimate emails according to particular criteria customized to the type of institutional sender (bank, online vendor) are both possible and critical. To focus this kind of effort, the customization of evaluation criteria to specific institutions (e.g. Chase, Bank of America) could be accomplished for a limited set of the most popular, visited, or prominent online institutions.

11.5 Attack Models

Altusys has preliminarily modeled the process by which the Phisher pursues the attack in order to develop defense mechanisms that target the attack logic/process:

Step 1

Phisher chooses individual or set of individuals to target. These could be employees of a certain company or individuals who post comments on a particular set of websites or social networking sites.

Example: Select all individuals who posted questions, answers, and comments to the Quora.com pages on corporate finance topics and select their followers [several hundred people total]. Requires the Phisher to create a Quora account, sign on, search for and navigate to the corporate finance topic page, then accumulate Quora names/profile information for all individuals who are following or posting on the page as well as their followers.

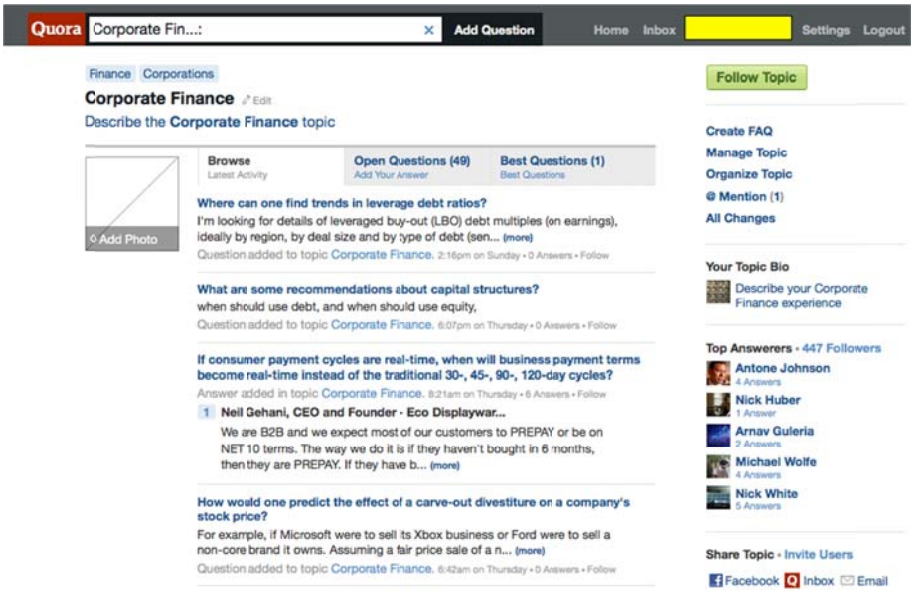


Figure 10 Quora.com Corporate Finance topic page

Step 2

Track patterns in posting and comments between users of the site in order to map the most common communication flows and connections.

Who is in contact with whom, when, and how often?

See Figure 11

Step 3

Determine most frequent topics listed in order to craft a message that is relevant and interesting for the target recipients.

Who posts or comments about which topics?

How are different topics connected via posters and commenters?

Are there patterns in the sequences of posting and navigation? (not addressed in graph below)

See Figure 12

Step 4

Analyze the message style and content patterns by author/topic

Who has a history of sending links? Are links more frequent for particular topics?

See Figure 12

What are the vocabulary sets, phrases, or sub-topics that reoccur in the communication? (Example report generated for Quora topic page on Corporate Finance using text analysis software)

See Figure 13

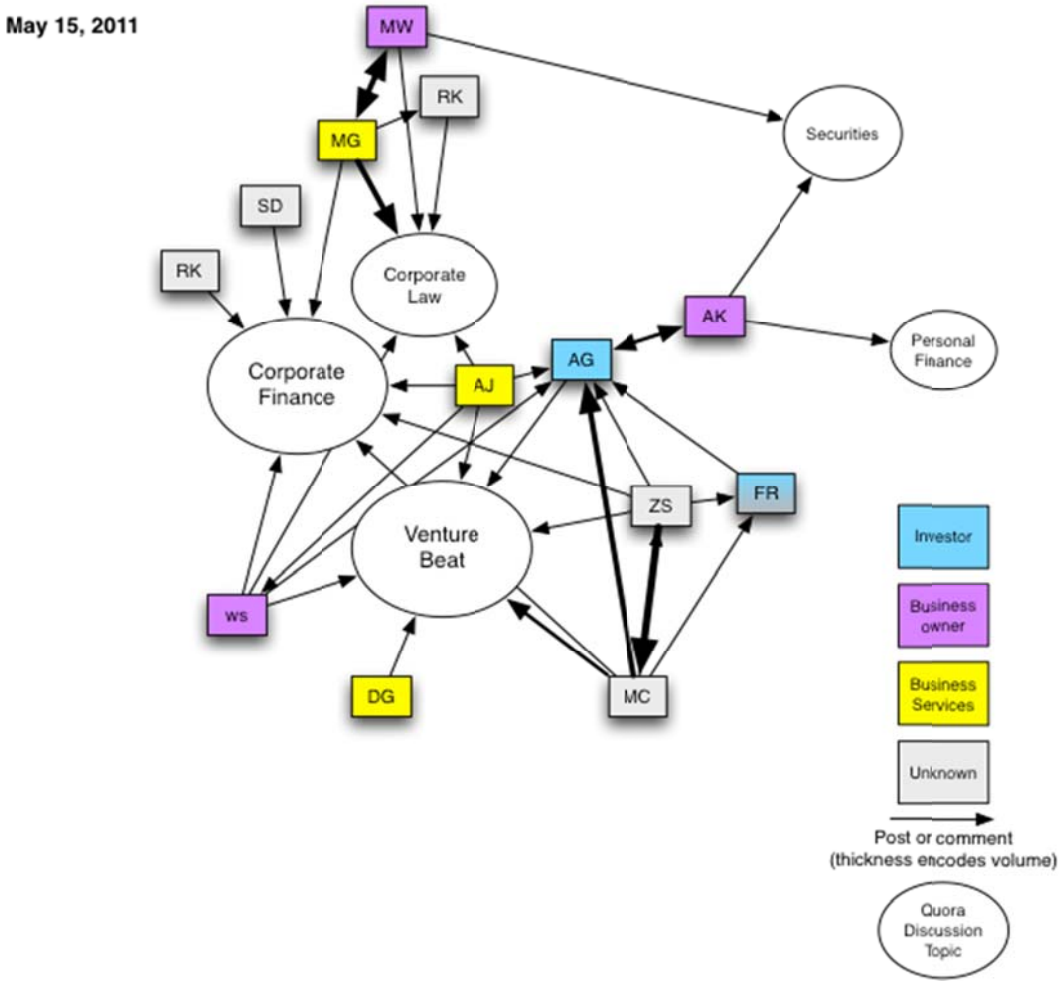


Figure 11 Patterns in posting and comments

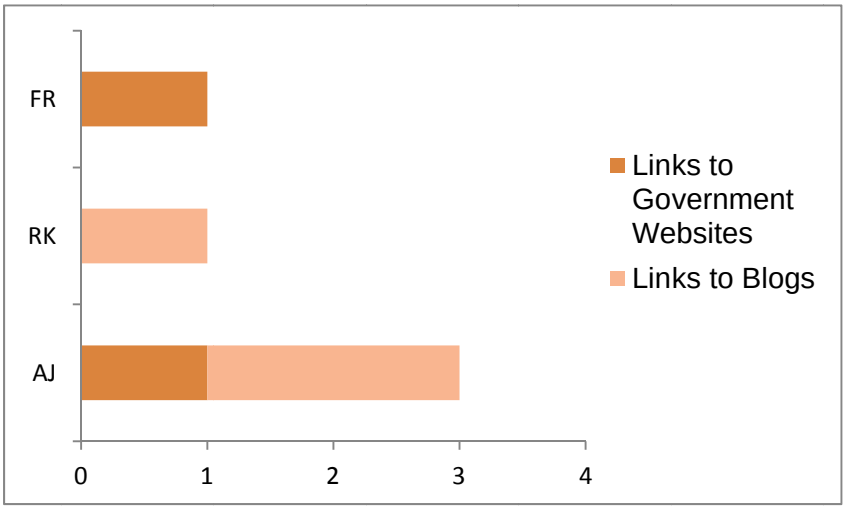


Figure 12 Quantity and Type of Links Posted, by User

- **Corporate Finance**
  - Initial Public Offerings
    - IPO and M&A Rumors
  - Mergers & Acquisitions
    - 8 more ...
  - Private Equity
    - Angel Investing
      - 6 more ...
    - Growth Capital
    - Distressed Investments
    - Mezzanine Capital
    - Secondary Investments
    - Carried Interest
  - Corporate Structures
    - C-Corps
    - S Corporation
    - LLCs
    - Partnerships
    - Sole Proprietorships
    - Shareholders
      - 1 more ...
    - Delaware Corporations
  - Valuations
    - 409A Valuations
    - Earnings Before Interest, Taxes, Depreciation & Amortization
    - Mark to Market Valuation
    - How Much Is X Worth?
    - Appraisals
      - 1 more ...
    - Domain Appraisal
    - Market Capitalization
  - Leverage & Borrowing
    - Leveraged Buy-Outs
    - Deleveraging
    - Interest Rates

Expression	Expression count	Frequency	Prominence
in cash	5	0.2%	55.3
has in	5	0.2%	55.3
it has	5	0.2%	55.4
billion it	5	0.2%	55.4
billion	5	0.2%	55.5
the	5	0.2%	55.6
some of	5	0.2%	55.7
use some	5	0.2%	55.7
or use	5	0.2%	55.7
spend or	5	0.2%	55.8
apple spend	5	0.2%	55.8
should apple	5	0.2%	55.9
how should	5	0.2%	55.9
more corporate	5	0.2%	74.6
instead of	5	0.2%	77.9
marc bodnick	4	0.2%	54.2
view	4	0.2%	54.5
retailer corporate	4	0.2%	56
ecommerce retailer	4	0.2%	56
an ecommerce	4	0.2%	56.1
retailer an	4	0.2%	56.1
a retailer	4	0.2%	56.2
about a	4	0.2%	56.2
what about	4	0.2%	56.3
get what	4	0.2%	56.3

Figure 13 Text analysis of Quora topic list

Figure 13shows on left: Quora Topic Page Sub-topic list, and on right: Excerpt from Text Analysis of Quora Topic Page Most Frequent Phrases, excluding phrases/words that are part of the Quora page template (Analysis was done at http://textalyser.net)

Other key metrics about page content and style:

Average sentence length: 7.78 words (range: 1-46)

Word length: 83% were 7 characters or less

Posted message length: Range 1-45 lines, Avg. 3-4 lines

Other frequent words/phrases: "Right now," "improvement," "cancel," "update," "system," "capital," "spend," "gains"

Step 5

Compose a fraudulent email that conforms to existing patterns of communication and content.

Determine a Sender: AJ (Frequent poster, commenter linked to numerous members with history of posting links) (Impersonate him either by hacking his account or creating a new account that impersonates him)

Analyze Sender Message and Link Introduction Style (in Quora postings)

Determine recipients: All posters, commenters and their followers (Maximum reach)

Choose topic: News, Technology M&A Activity (Maps to Quora sub-topics, AJs exhibited areas of interest, and areas of interest/focus expressed in network members' postings)

Craft message and deceptive description of Phishing link:

From: Andy Johnson on Quora

To: Quora Corporate Finance community

Subject: Technology M&A Activity News and Analysis

For those of you in Quora Finance, Venture, and Investment communities who are interested in learning about more new developments in the technology and social media sectors ***see my new blog at*** www.andyjohnson.blogs.com. I will be providing daily news, insights, and updates.

AJ

11.6 HPPS Requirements

11.6.1 Functional Requirements

(Including customer and System requirements.)

11.6.1.1 HPPS Email Analyzer (EA)

1. HPSS EA must receive each incoming email to be analyzed
2. HPSS EA may receive an incoming email from a designated mail server.
3. HPSS EA may receive a selected set of emails stored in a file or Outlook mailbox.
4. HPSS EA may receive an incoming email from a spam analyzer which removes SPAM from the email stream.
5. For each email, HPSS EA must generate a psycho-social profile. The psycho-social profile may include author's intention. The psycho-social profile includes TBD.
6. For each email, HPSS EA must generate a structural labeling. The structure labeling includes TBD.
7. The profile includes statistical weighting of labels and attributes. The statistical classifier may include weighted statistical filters supporting complex non-linear feature combinations for structural feature analysis, Class-based Latent Dirichlet Allocation model for analyzing psycho-social features as well as cascading classifier techniques to minimize computational effort during classification.
8. The profile must include a unique identifier of the email and may include a copy of the email. The copy of the email may be anonymized, i.e., removing sender and receiver information.
9. The email profile may be stored in a database.
10. The email profile should be used to trigger an alert to the user if the profile indicates the email is likely a phishing email.
11. HPPS EA algorithms should be adaptive.

11.6.1.2 HPPS Phishing Web Site Analyzer (PWSA)

12. HPPS PWSA acts as a web site client.
13. HPSS PWSA connects to a url found in an email. HPSS PWSA uses http or https to do this.
14. HPSS PWSA may be configured to emulate the http signature of a specific web browser.
15. HPSS PWSA must not pass any identifiable user information or location information to the site.
16. HPSS PWSA must store the response to each GET request in a database. Each stored entry must include the complete URL, a copy of the request, and complete timestamp.
17. HPSS PWSA may automatically GET other objects referenced in the response to the GET url. All objects retrieved must be stored in the database, with a complete URL, a copy of the request, and a complete timestamp.

18. HPSS PWSA may truncate stream objects or any objects exceeding a specific size.
19. HPSS PWSA may crawl all or portions of the web site. All objects retrieved must be stored in the database, with a complete URL, a copy of the request, and a complete timestamp.
20. HPSS PWSA may extract any embedded URL found in a web page retrieved from the site and store that URL in the database.
21. HPSS PWSA may crawl to other websites whose URLs have been found in content retrieved from the original website.
22. HPPS PWSA algorithms should be adaptive.
23. For each web page retrieved, HPPS PWSA must create a profile of phish web pages appearances built with fuzzy hashing techniques to detect such Web pages on the client side.
24. The HPPS PWSA must inspect the fraudulent Web page for defining content and common characteristics of many phishing campaigns and creates a phish web page profile.
25. HPPS should utilize input from the IDS and correlates that into analysis to help determine if the Web site is forged or if the mail server is compromised.
26. HPPS does not rely on any black lists. It develops email as well as phishing web page profile automatically and distributes them and hence it is able to cope up with fast flux attacks.
27. HPPS ensures near-zero false positives by having detection methods at every stage of the attack. If the user chooses to ignore the warning at one stage then the user is warned at the next stage of the attack. Finally, if the user completely disregards the HPPS's warnings then HPPS acts in the active defense mode by submitting misinformation to the Phisher and thereby defending against the attack. HPPS is adaptive and ensure low latency in detection by automatically developing and propagating new phish email and web page profiles to other HPPSs and hardening other systems.

11.6.1.3 HPPS Anti-Phishing Intervention (AI)

28. HPPS AI is used to inject misinformation to the Phisher along with user's real credentials to actively disrupt phishing activity.
29. HPPS AI must be able to transparently intervene between a web browser and phishing website for an http connection
30. HPPS AI should be able to transparently intervene between a web browser and phishing website for an https connection.
31. HPPS AI may act as a proxy server for multiple web browser clients.
32. HPPS AI should record each session into a database. The session record should be associated with a specific phishing email.

11.6.1.4 HPPS Distributed Coordination (DC)

33. HPSS DC is the means by which separate HPPS systems share email and phishing web site profiles. This allows each DC to correlate mass phishing attacks, increases probability of belief for an identification of a phishing email, and may accelerate identification of phishing attack emails.
34. HPPS DC should use a wide-area secure publish/subscribe mechanism.
35. HPSS DC may use the Altusys secure overlay with wide-area publish/subscribe capability.
36. TBD: should there be a single topic for all object (e.g., HPSS) or should there be specific topics, such as Language specific phishing (English, Chinese, …), Source specific phishing, Type specific phishing (spear phishing, link manipulation, real-estate scam, adult content, fake lottery, chain letters, personal finance, pharmaceutical or viagra, stock pumping, nigerian letters or 419 advance fee fraud, degrees, casino, weight loss, etc.)
37. Each email and web site profile should be published to all subscribers as soon as it is identified as a new phishing attack.
38. HPSS DC node should determine if it has already received an equivalent notification from another node, in which cast it should not publish it to the subscription network.

11.6.1.5 HPPS Administration Interface (ADM)

39. HPSS ADM must be used to configure HPSS EA.
40. HPPS ADM must be used to configure HPSS PWSA
41. HPPS ADM must be used to configure HPSS AI
42. HPPS ADM must be used to configure HPSS DC
43. HPSS ADM is accessible through a web interface.
44. HPPS EA email profile may be viewable through an HPSS ADM.
45. The EA email trigger conditions for alerting a user should be configurable through an HPSS administration interface.
46. The extent of HPSS PWSA crawling within a given website should be configurable
47. The extent of HPSS PWSA crawling to sites linked from the orginal web site should be configurable.

11.6.2 HPSS Test Data

48. HPSS Test Data is a set of emails for evaluating HPPS algorithms.
49. HPSS Test Data should include recent phishing emails, since lifetime of phishing sites is short
50. HPSS Test Data should include phishing emails using social contacts.
51. HPSS Test Data should include known SPAM categories (List of categories goes here).

11.6.3 Strategic requirements

(Including legacy interfaces and replacements, competition, benchmarks, and market window.)

52. HPPS should use existing software components where available
53. HPPS should work with existing email spam filters and analyzers
54. HPSS should use SABIA secure overlay for HPSS DC.

11.6.4 Architectural requirements

55. HPSS must be interoperable with email servers and web servers using standard protocols including SMTP and http.
56. HPSS should use COTS database, for example MySQL.
57. HPSS AI should be architected as a proxy
58. HPSS ADM should be implemented using a Tomcat server.
59. HPSS may use Java Mail API for integration with mail servers
60. HPSS should use open source components where possible, such as linguistic analysis.

11.6.5 Security requirements

(Including access control, authentication, authorization)

61. HPPS ADM should require authentication credentials for login
62. HPPS ADM may have different user roles
63. A user should be able to see the email and phishing site profile for any email sent to them which has been identified by HPPS as a phishing email
64. A user must not be able to see email and/or phishing site profiles for emails sent to other users.

11.6.6 Performance requirements

(Including expected application environments and other load factors for the product)

65. HPPS should demonstrate the ability to respond to a fast flux attack.
66. HPSS should have false positive rate of < 10% for the test data set
67. HPSS should have false negative rate of < 10% for the test data set

11.6.7 Scalability requirements

(Including future applications, demand growth, and similar factors over the life of the product)

68. HPPS DC should scale to thousands of nodes with hundreds of simultaneous publishers
69. HPPS EA should be able to handle thousands of users simultaneously

11.6.8 Testability requirements

(Including regression test, unit test, system test; test data generation, test execution process, and test validation method)

70. HPSS prototype must be able to process emails stored in a local folder, either txt or format used by MS Outlook.
71. HPSS prototype should be able to process emails via connecting to email server with specified credentials, such as hpps@altusystems.com
72. HPSS development environment should include JUnit testing capability.

73. HPSS development environment may include nightly build automated testing.
74. The development plan must include a labeled dataset of emails
75. Live testing sessions must be recorded for post-analysis

11.6.9 Documentation and Training

(Including materials preparation, interchange with other organizations, user community needs)

76. HPSS software must include documentation for requirements and design.
77. HPPS source code must be structured into packages or modules
78. HPPS source code must be commented

11.6.10 System Admin

(Document installation procedures and needs for maintaining operations environment)

79. HPSS final version must include installation, configuration, and operation instructions for Windows environment.
80. HPSS final version must include installation, configuration, and operation instructions for any other environment(s) which are supported

11.6.11 Error Handling

(Acceptable behavior under anomalous conditions, error message output, needed alarm/alerts)

81. HPPS errors should be logged at an appropriate severity level
82. HPPS run-time errors should be caught without disrupting operation of the application

11.6.12 Availability (Fault Tolerance)

(If needed, specify needs for continuous operation, degradation of performance under load, requirements for restart or automatic startup at host boot, etc. Are system crashes acceptable?)

83. There are no availability requirements in this version

11.6.13 Third Party Software

(Requirements for OS, db, and other support or data sources when known.)

84. HPPS should run on current Windows platforms such as WinXP, Windows 7, Vista
85. HPPS may run on Linux

11.6.14 Century Compliance and Internationalization

(How are two-digit years interpreted, if applicable.)

86. HPPS must be century compliant
87. HPPS should be internationalized w.r.t. to processing email in different languages.
88. HPPS may be used in non-English email environments.